Неявная кросс-валидация для выбора подмножества информативных признаков в задаче обучения распознаванию образов по методу опорных векторов

Черноусова Елена Олеговна
Левдик Павел Владимирович
Московский физико-технический институт

Моттль Вадим Вячеславович
Вычислительный центр РАН

Уиндридж, Дэвид
Университет Суррей, Великобритания

Типовая задача восстановления закономерностей в множествах объектов реального мира

Типовая задача восстановления закономерностей в множествах объектов реального мира

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Типовая задача восстановления закономерностей в множествах объектов реального мира

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Типовая задача восстановления закономерностей в множествах объектов реального мира

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$

Типовая задача восстановления закономерностей в множествах объектов реального мира

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$

Желание наблюдателя:
Иметь инструмент оценивания скрытой характеристики для реальных объектов
$\hat{y}(x) : \mathbb{X} \to \mathbb{Y}; \quad \hat{y}(x) \neq y(x)$ – ошибка.

Типовая задача восстановления закономерностей в множествах объектов реального мира

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$

Желание наблюдателя:
Иметь инструмент оценивания скрытой характеристики для реальных объектов
$\hat{y}(x) : \mathbb{X} \to \mathbb{Y}$; $\hat{y}(x) \neq y(x)$ – ошибка.

Обучение по прецедентам:

Подмножество наблюдаемых объектов, для которых измерены значения обеих характеристик $\Omega^* \subset \Omega$: $\left\{ \left(x(\omega_j), y(\omega_j) \right) = (x_j, y_j),\ j = 1, \ldots, N \right\}$.

Типовая задача восстановления закономерностей в множествах объектов реального мира

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$

Желание наблюдателя:
Иметь инструмент оценивания скрытой характеристики для реальных объектов
$\hat{y}(x) : \mathbb{X} \to \mathbb{Y}$; $\hat{y}(x) \neq y(x)$ – ошибка.

Обучение по прецедентам:

Подмножество наблюдаемых объектов, для которых измерены значения обеих характеристик $\Omega^* \subset \Omega : \left\{ \left(x(\omega_j), y(\omega_j) \right) = (x_j, y_j),\ j = 1, \ldots, N \right\}$.

Задача: Продолжить функцию на все множество Ω, так чтобы можно было в дальнейшем оценивать значение рассматриваемой характеристики $\hat{y}(\omega) = \hat{y}\left(x(\omega) \right)$ для новых объектов $\omega \in \Omega \setminus \Omega^*$.

Типовая задача восстановления закономерностей в множествах объектов реального мира

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$

Желание наблюдателя:
Иметь инструмент оценивания скрытой характеристики для реальных объектов
$\hat{y}(x) : \mathbb{X} \to \mathbb{Y}$; $\hat{y}(x) \neq y(x)$ – ошибка.

Обучение по прецедентам:

Подмножество наблюдаемых объектов, для которых измерены значения обеих характеристик $\Omega^* \subset \Omega$: $\left\{\left(x(\omega_j), y(\omega_j)\right) = (x_j, y_j),\ j = 1,\ldots,N\right\}$.

Простейшие случаи:
Задача распознавания образов
$\mathbb{Y} = \{y_1,\ldots,y_m\}$ – конечное неупорядоченное множество; в частности $\mathbb{Y} = \{-1, 1\}$.

Задача восстановления числовой зависимости
$\mathbb{Y} = \mathbb{R}$ – множество действительных чисел.

Вероятностная интерпретация задачи восстановления зависимости по обучающей совокупности

Вероятностная интерпретация задачи восстановления зависимости по обучающей совокупности

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Плотность распределения наблюдателю неизвестна $f^*(x, y) \geq 0$, $\int\limits_{\mathbb{Y}} \int\limits_{\mathbb{X}} f^*(x, y) dx dy = 1$.

Вероятностная интерпретация задачи восстановления зависимости по обучающей совокупности

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Плотность распределения наблюдателю неизвестна $f^*(x, y) \geq 0$, $\int\limits_{\mathbb{Y}} \int\limits_{\mathbb{X}} f^*(x, y) dx dy = 1$.

Тем не менее, наблюдатель вынужден выбрать оценку $\hat{y}(x, \mathbf{a})$, $\mathbf{a} \in \mathbb{R}^n$, предполагая некоторую функцию потерь $Loss\left(y, \hat{y}(x, \mathbf{a})\right) = q(y, x, \mathbf{a})$.

Вероятностная интерпретация задачи восстановления зависимости по обучающей совокупности

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Плотность распределения наблюдателю неизвестна $f^*(x, y) \geq 0$, $\int\limits_{\mathbb{Y}} \int\limits_{\mathbb{X}} f^*(x, y) dx dy = 1.$

Тем не менее, наблюдатель вынужден выбрать оценку $\hat{y}(x, \mathbf{a})$, $\mathbf{a} \in \mathbb{R}^n$, предполагая некоторую функцию потерь $Loss\left(y, \hat{y}(x, \mathbf{a})\right) = q(y, x, \mathbf{a})$.

Средний риск ошибки: $r_{av}(\mathbf{a}) = r_{av}\left[\hat{y}(\bullet, \mathbf{a})\right] = \int\limits_{\mathbb{X}} \int\limits_{\mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx$

Вероятностная интерпретация задачи восстановления зависимости по обучающей совокупности

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Плотность распределения наблюдателю неизвестна $f^*(x, y) \geq 0$, $\int\limits_{\mathbb{Y}} \int\limits_{\mathbb{X}} f^*(x, y) dx dy = 1.$

Тем не менее, наблюдатель вынужден выбрать оценку $\hat{y}(x, \mathbf{a})$, $\mathbf{a} \in \mathbb{R}^n$, предполагая некоторую функцию потерь $Loss\left(y, \hat{y}(x, \mathbf{a})\right) = q(y, x, \mathbf{a})$.

Средний риск ошибки: $$r_{av}(\mathbf{a}) = r_{av}\left[\hat{y}(\bullet, \mathbf{a})\right] = \int\limits_{\mathbb{X}} \int\limits_{\mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx$$

Естественный способ выбора решающего правила – минимизация среднего риска

$$r_{av}(\mathbf{a}) = r_{av}\left[\hat{y}(\bullet, \mathbf{a})\right] = \int\limits_{\mathbb{X}} \int\limits_{\mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx \to \min(\mathbf{a})$$

Вероятностная интерпретация задачи восстановления зависимости по обучающей совокупности

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Плотность распределения наблюдателю неизвестна $f^*(x, y) \geq 0$, $\int\limits_{\mathbb{Y}} \int\limits_{\mathbb{X}} f^*(x, y) dx dy = 1.$

Тем не менее, наблюдатель вынужден выбрать оценку $\hat{y}(x, \mathbf{a})$, $\mathbf{a} \in \mathbb{R}^n$, предполагая некоторую функцию потерь $Loss\left(y, \hat{y}(x, \mathbf{a})\right) = q(y, x, \mathbf{a})$.

Средний риск ошибки: $$r_{av}(\mathbf{a}) = r_{av}\left[\hat{y}(\bullet, \mathbf{a})\right] = \int\limits_{\mathbb{X}} \int\limits_{\mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx$$

Естественный способ выбора решающего правила – минимизация среднего риска

$$r_{av}(\mathbf{a}) = r_{av}\left[\hat{y}(\bullet, \mathbf{a})\right] = \int\limits_{\mathbb{X}} \int\limits_{\mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx \to \min(\mathbf{a})$$

Проклятие неопределенности! Плотность $f^*(x, y)$ неизвестна наблюдателю!

Вероятностная интерпретация задачи восстановления зависимости по обучающей совокупности

Некоторое множество реально существующих объектов $\omega \in \Omega$.

С каждым объектом связана пара характеристик $\begin{cases} x(\omega) \in \mathbb{X} - \text{наблюдаемая} \\ y(\omega) \in \mathbb{Y} - \text{скрытая} \end{cases}$

Природа случайным образом многократно и независимо выбирает объект $\omega \in \Omega$, то есть пару $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Плотность распределения наблюдателю неизвестна $f^*(x, y) \geq 0$, $\int\limits_{\mathbb{Y}} \int\limits_{\mathbb{X}} f^*(x, y) dx dy = 1.$

Тем не менее, наблюдатель вынужден выбрать оценку $\hat{y}(x, \mathbf{a})$, $\mathbf{a} \in \mathbb{R}^n$, предполагая некоторую функцию потерь $Loss\left(y, \hat{y}(x, \mathbf{a})\right) = q(y, x, \mathbf{a})$.

Средний риск ошибки: $$r_{av}(\mathbf{a}) = r_{av}\left[\hat{y}(\bullet, \mathbf{a})\right] = \int\limits_{\mathbb{X}} \int\limits_{\mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx$$

Естественный способ выбора решающего правила – минимизация среднего риска

$$r_{av}(\mathbf{a}) = r_{av}\left[\hat{y}(\bullet, \mathbf{a})\right] = \int\limits_{\mathbb{X}} \int\limits_{\mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx \to \min(\mathbf{a})$$

Проклятие неопределенности! Плотность $f^*(x, y)$ неизвестна наблюдателю!
Стандартный выход – минимизация регуляризованного эмпирического риска.

Выбор решающего правила по критерию минимума регуляризованного эмпирического риска

Выбор решающего правила по критерию минимума регуляризованного эмпирического риска

Два «островка определенности», доступные наблюдателю.

1) Обучающая совокупность $(\mathbf{x},\mathbf{y}) = \{(x_j, y_j),\ j = 1,\ldots,N\}$

   Эмпирический риск выбора решающего правила $Q(\mathbf{y},\mathbf{x},\mathbf{a}) = \left(\frac{1}{N}\right)\sum_{j=1}^{N} q(y_j, x_j, \mathbf{a})$ –

   несмещенная оценка сред среднего риска по обучающей совокупности
   Желание наблюдателя – минимизировать хотя бы эмпирический риск
   $Q(\mathbf{y},\mathbf{x},\mathbf{a}) \to \min(\mathbf{a})$.

2) Априорные предпочтения относительно параметра решающего правила
   $V(\mathbf{a}, C) \to \min(\mathbf{a})$, C – скалярный либо векторный «структурный» параметр.

Выбор решающего правила по критерию минимума регуляризованного эмпирического риска

Два «островка определенности», доступные наблюдателю.

1) Обучающая совокупность $(\mathbf{x},\mathbf{y}) = \{(x_j, y_j),\ j = 1,\ldots,N\}$

   Эмпирический риск выбора решающего правила $Q(\mathbf{y},\mathbf{x},\mathbf{a}) = \left(\frac{1}{N}\right)\sum_{j=1}^{N} q(y_j, x_j, \mathbf{a})$ –

   несмещенная оценка сред среднего риска по обучающей совокупности

   Желание наблюдателя – минимизировать хотя бы эмпирический риск $Q(\mathbf{y},\mathbf{x},\mathbf{a}) \to \min(\mathbf{a})$.

2) Априорные предпочтения относительно параметра решающего правила $V(\mathbf{a}, C) \to \min(\mathbf{a})$, C – скалярный либо векторный «структурный» параметр

Компромисс – минимизация регуляризованного эмпирического риска

Критерий обучения: $\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x}) = \arg\min_{\mathbf{a}} \{V(\mathbf{a}, C) + Q(\mathbf{y},\mathbf{x},\mathbf{a})\}$

Выбор решающего правила по критерию минимума регуляризованного эмпирического риска

Два «островка определенности», доступные наблюдателю.

1) Обучающая совокупность $(\mathbf{x},\mathbf{y}) = \{(x_j, y_j),\ j = 1,\ldots,N\}$

   Эмпирический риск выбора решающего правила $Q(\mathbf{y},\mathbf{x},\mathbf{a}) = \left(\frac{1}{N}\right)\sum_{j=1}^{N} q(y_j, x_j, \mathbf{a})$ –

   несмещенная оценка сред среднего риска по обучающей совокупности

   Желание наблюдателя – минимизировать хотя бы эмпирический риск $Q(\mathbf{y},\mathbf{x},\mathbf{a}) \to \min(\mathbf{a})$.

2) Априорные предпочтения относительно параметра решающего правила $V(\mathbf{a}, C) \to \min(\mathbf{a})$, C – скалярный либо векторный «структурный» параметр

Компромисс – минимизация регуляризованного эмпирического риска

Критерий обучения: $\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x}) = \arg\min_{\mathbf{a}} \{V(\mathbf{a}, C) + Q(\mathbf{y},\mathbf{x},\mathbf{a})\}$

Как выбрать структурный параметр C?

Выбор решающего правила по критерию минимума регуляризованного эмпирического риска

Два «островка определенности», доступные наблюдателю.

1) Обучающая совокупность $(\mathbf{x},\mathbf{y}) = \{(x_j, y_j),\ j = 1,\ldots,N\}$

Эмпирический риск выбора решающего правила $Q(\mathbf{y},\mathbf{x},\mathbf{a}) = \left(\frac{1}{N}\right)\sum_{j=1}^{N} q(y_j, x_j, \mathbf{a})$ –

несмещенная оценка сред среднего риска по обучающей совокупности

Желание наблюдателя – минимизировать хотя бы эмпирический риск $Q(\mathbf{y},\mathbf{x},\mathbf{a}) \to \min(\mathbf{a})$.

2) Априорные предпочтения относительно параметра решающего правила $V(\mathbf{a}, C) \to \min(\mathbf{a})$, C – скалярный либо векторный «структурный» параметр

Компромисс – минимизация регуляризованного эмпирического риска

Критерий обучения: $\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x}) = \arg\min_{\mathbf{a}} \{V(\mathbf{a}, C) + Q(\mathbf{y},\mathbf{x},\mathbf{a})\}$

Как выбрать структурный параметр C?

Выбрать его по критерию минимума среднего риска невозможно

$$\int_{\mathbb{Y}}\int_{\mathbb{X}} Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) F^*(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \to \min(C), \quad \underbrace{F^*(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{N} f^*(x_j, y_j)}$$

неизвестное совместное распределение случайной обучающей выборки

Выбор решающего правила по критерию минимума регуляризованного эмпирического риска

Два «островка определенности», доступные наблюдателю.

1) Обучающая совокупность $(\mathbf{x},\mathbf{y}) = \{(x_j, y_j),\ j = 1,\ldots,N\}$

   Эмпирический риск выбора решающего правила $Q(\mathbf{y},\mathbf{x},\mathbf{a}) = \left(\frac{1}{N}\right)\sum_{j=1}^{N} q(y_j, x_j, \mathbf{a})$ –

   несмещенная оценка сред среднего риска по обучающей совокупности

   Желание наблюдателя – минимизировать хотя бы эмпирический риск $Q(\mathbf{y},\mathbf{x},\mathbf{a}) \to \min(\mathbf{a})$.

2) Априорные предпочтения относительно параметра решающего правила $V(\mathbf{a}, C) \to \min(\mathbf{a})$, C – скалярный либо векторный «структурный» параметр

Компромисс – минимизация регуляризованного эмпирического риска

Критерий обучения: $\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x}) = \arg\min_{\mathbf{a}} \{V(\mathbf{a}, C) + Q(\mathbf{y},\mathbf{x},\mathbf{a})\}$

Как выбрать структурный параметр C?

Выбрать его по критерию минимума среднего риска невозможно

$$\int_{\mathbb{Y}}\int_{\mathbb{X}} Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) F^*(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \to \min(C), \quad \underbrace{F^*(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{N} f^*(x_j, y_j)}$$

неизвестное совместное распределение случайной обучающей выборки

В этом заключается одна из основных проблем теории обучения.
Именно эта проблема составляет предмет данного доклада.

Предпосылки исследования: Методы выбора структурного параметра обучения по единственной обучающей совокупности

- **Минимизация структурного риска** (В.Н. Вапник, А.Я. Червоненкис, 1974). ***Бесперебо́рный метод.*** Основан на понятии VC-dimension и применим только к критериям обучения распознаванию образов.
- **Информационный критерий Акаике** (Hirotugu Akaike, 1974). ***Бесперебо́рный метод.*** Основан на априорной упорядоченности элементов вектора параметров $\mathbf{a}=(a_1,...,a_\lambda,a_{\lambda+1}=0,...,a_n=0)$ и применим только к квадратичным критериям обучения.
- **Скользящий контроль** (М.Н. Вайнгцвайг, 1971). Универсальный ***переборный метод***. Последовательное выделение одного объекта обучающей совокупности для контроля ошибки при обучении по остальным объектам. ***Высокая вычислительная сложность***.
- **Кросс-валидация – Jackknife** (Bradley Efron, 1982). Более общий ***переборный метод***. Основан на многократном делении выборки на обучающую и контрольную части. ***Вычислительная сложность может быть еще больше***.

Задачи данного исследования

- Показать, что идея информационного критерия Акаике для выбора структурного параметра основана на ***принципе неявной кросс-валидации***.
- Разработать ***бесперебо́рный критерий кросс-валидации*** для квадратичной задачи оценивания линейной регрессии, в котором классический критерий Акаике являлся бы частным случаем.
- Разработать ***бесперебо́рный критерий кросс-валидации*** для задачи обучения распознаванию образов по методу опорных векторов – Support Vector Machine, SVM.

Принцип неявной кросс-валидации: Первая эвристики наблюдателя

Принцип неявной кросс-валидации: Первая эвристики наблюдателя

Еще раз неизвестная совместная плотность распределения наблюдаемой и скрытой характеристик объекта	$\underbrace{F^*(\mathbf{x},\mathbf{y})}_{\textit{выборка в целом}} = \prod_{j=1}^{N} \underbrace{f^*(x_j, y_j)}_{\textit{отдельные случайные объекты}}$

Принцип неявной кросс-валидации: Первая эвристики наблюдателя

Еще раз неизвестная совместная плотность распределения наблюдаемой и скрытой характеристик объекта	$\underbrace{F^*(\mathbf{x},\mathbf{y})}_{\textit{выборка в целом}} = \prod_{j=1}^{N} \underbrace{f^*(x_j, y_j)}_{\textit{отдельные случайные объекты}}$

Представим ее в виде произведения маргинальной плотности распределения наблюдаемой характеристики и условной плотности распределения ненаблюдаемой характеристики:

$$f^*(x,y) = \varphi^*(y \mid x) g^*(x)$$

$$F^*(\mathbf{x},\mathbf{y}) = \underbrace{\prod_{j=1}^{N} \varphi^*(y_j \mid x_j)}_{\Phi^*(\mathbf{y}|\mathbf{x})} \underbrace{\prod_{j=1}^{N} g^*(x_j)}_{G^*(\mathbf{x})} = \boxed{\Phi^*(\mathbf{y} \mid \mathbf{x})} G^*(\mathbf{x})$$

Принцип неявной кросс-валидации: Первая эвристики наблюдателя

Еще раз неизвестная совместная плотность распределения наблюдаемой и скрытой характеристик объекта	$\underbrace{F^*(\mathbf{x},\mathbf{y})}_{\text{выборка в целом}} = \prod_{j=1}^{N} \underbrace{f^*(x_j, y_j)}_{\text{отдельные случайные объекты}}$

Представим ее в виде произведения маргинальной плотности распределения наблюдаемой характеристики и условной плотности распределения ненаблюдаемой характеристики:

$$f^*(x,y) = \varphi^*(y \mid x) g^*(x)$$

$$F^*(\mathbf{x},\mathbf{y}) = \underbrace{\prod_{j=1}^{N} \varphi^*(y_j \mid x_j)}_{\Phi^*(\mathbf{y}\mid\mathbf{x})} \underbrace{\prod_{j=1}^{N} g^*(x_j)}_{G^*(\mathbf{x})} = \boxed{\Phi^*(\mathbf{y} \mid \mathbf{x})} G^*(\mathbf{x})$$

Первая эвристика наблюдателя: Предположение об условном распределении скрытой характеристики

Неизвестная плотность распределения представляется в виде **неизвестной смеси известных распределений**	$\Phi^*(\mathbf{y}\mid\mathbf{x}) = \int_{\mathbb{R}^n} \underbrace{\Phi(\mathbf{y} \mid \mathbf{x},\mathbf{a})}_{\substack{\text{известное параметрическое} \\ \text{семейство}}} \underbrace{\Psi^*(\mathbf{a})}_{\substack{\text{неизвестная} \\ \text{смесь}}} d\mathbf{a}$

Принцип неявной кросс-валидации: Первая эвристики наблюдателя

Еще раз неизвестная совместная плотность распределения наблюдаемой и скрытой характеристик объекта	$\underbrace{F^*(\mathbf{x},\mathbf{y})}_{\text{выборка в целом}} = \prod_{j=1}^{N} \underbrace{f^*(x_j,y_j)}_{\text{отдельные случайные объекты}}$

Представим ее в виде произведения маргинальной плотности распределения наблюдаемой характеристики и условной плотности распределения ненаблюдаемой характеристики:

$$f^*(x,y) = \varphi^*(y \mid x) g^*(x)$$

$$F^*(\mathbf{x},\mathbf{y}) = \underbrace{\prod_{j=1}^{N} \varphi^*(y_j \mid x_j)}_{\Phi^*(\mathbf{y}\mid\mathbf{x})} \underbrace{\prod_{j=1}^{N} g^*(x_j)}_{G^*(\mathbf{x})} = \boxed{\Phi^*(\mathbf{y} \mid \mathbf{x})} G^*(\mathbf{x})$$

Первая эвристика наблюдателя: Предположение об условном распределении скрытой характеристики

Неизвестная плотность распределения представляется в виде **неизвестной смеси известных распределений**	$\Phi^*(\mathbf{y}\mid\mathbf{x}) = \int_{\mathbb{R}^n} \underbrace{\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})}_{\substack{\text{известное параметрическое} \\ \text{семейство}}} \underbrace{\Psi^*(\mathbf{a})}_{\substack{\text{неизвестная} \\ \text{смесь}}} d\mathbf{a}$
Принцип незлонамеренности природы. Наблюдатель правильно «угадал» функцию потерь и класс решающих правил. Природа чаще генерирует пары с низким значением штрафа.	$\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a}) \propto \exp\{-\underbrace{Q(\mathbf{y},\mathbf{x},\mathbf{a})}_{\substack{\text{принятая наблюдателем} \\ \text{функция потерь}}}\}$ *коэффициент не зависит от* $(\mathbf{x},\mathbf{a})$

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y} \mid \mathbf{x}, \mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,...,y_N)\in\mathbb{R}^N$ $\tilde{\mathbf{y}}=(\tilde{y}_1,...,\tilde{y}_N)\in\mathbb{R}^N$ Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y} \mid \mathbf{x}, \mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,...,y_N)\in\mathbb{R}^N$ $\tilde{\mathbf{y}}=(\tilde{y}_1,...,\tilde{y}_N)\in\mathbb{R}^N$ Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая
Мысленная перекрестная проверка на генеральной совокупности	Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения C вычислить оценку $\mathbf{a}$ по обучающей совокупности и подставить в функцию потерь для контрольной совокупности: $\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x}) \to Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)$ – потери на мысленном контроле

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,\ldots,y_N)\in\mathbb{R}^N$ $\tilde{\mathbf{y}}=(\tilde{y}_1,\ldots,\tilde{y}_N)\in\mathbb{R}^N$ Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая
Мысленная перекрестная проверка на генеральной совокупности	Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения C вычислить оценку $\mathbf{a}$ по обучающей совокупности и подставить в функцию потерь для контрольной совокупности: $\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\to Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)$ – потери на мысленном контроле
Идея скрытой кросс-валидации: минимум математического ожидания потерь	$\iiint Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x}=$ $\iiint Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)\Big\{\int\Phi(\tilde{\mathbf{y}}\mid\mathbf{x},\mathbf{a})\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})\,d\mathbf{a}\Big\}G^*(\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x}\to\min(C)$

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y} \mid \mathbf{x}, \mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,\ldots,y_N)\in\mathbb{R}^N$ $\tilde{\mathbf{y}}=(\tilde{y}_1,\ldots,\tilde{y}_N)\in\mathbb{R}^N$ Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая
Мысленная перекрестная проверка на генеральной совокупности	Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения C вычислить оценку $\mathbf{a}$ по обучающей совокупности и подставить в функцию потерь для контрольной совокупности: $\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x}) \to Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)$ – потери на мысленном контроле
Идея скрытой кросс-валидации: минимум математического ожидания потерь	$\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} =$ $\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)\left\{\int \Phi(\tilde{\mathbf{y}}\mid\mathbf{x},\mathbf{a})\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})\,d\mathbf{a}\right\}G^*(\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} \to \min(C)$

Согласно первой эвристике наблюдателя $Q(\mathbf{x},\mathbf{y},\mathbf{a}) = -\ln\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a}) + const$.

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,\ldots,y_N)\in\mathbb{R}^N$ $\tilde{\mathbf{y}}=(\tilde{y}_1,\ldots,\tilde{y}_N)\in\mathbb{R}^N$ Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая
Мысленная перекрестная проверка на генеральной совокупности	Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения C вычислить оценку $\mathbf{a}$ по обучающей совокупности и подставить в функцию потерь для контрольной совокупности: $\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\to Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)$ – потери на мысленном контроле
Идея скрытой кросс-валидации: минимум математического ожидания потерь	$\int\int\int Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x}=$ $\int\int\int Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)\Big\{\int\Phi(\tilde{\mathbf{y}}\mid\mathbf{x},\mathbf{a})\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})d\mathbf{a}\Big\}G^*(\mathbf{x})d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x}\to\min(C)$

Согласно первой эвристике наблюдателя $Q(\mathbf{x},\mathbf{y},\mathbf{a})=-\ln\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})+const$.

Идея скрытой кросс-валидации $\int\Big[\ln\Phi\big(\mathbf{y}|\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)\Big]\Phi(\mathbf{y}|\mathbf{x},\mathbf{a})dy\to\max$

есть максимизация информации Кульбака о распределении $\Phi(\mathbf{y}|\mathbf{x},\mathbf{a})$, содержащейся в его оценке по другой выборке $\Phi\big(\mathbf{y}|\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)$.

Поэтому критерии неявной кросс-валидации уместно называть информационными и рассматривать их как обобщение классической идеи Хиротугу Акаике.

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y} \mid \mathbf{x}, \mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,\ldots,y_N)\in\mathbb{R}^N$, $\tilde{\mathbf{y}}=(\tilde{y}_1,\ldots,\tilde{y}_N)\in\mathbb{R}^N$. Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая
Мысленная перекрестная проверка на генеральной совокупности	Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения C вычислить оценку $\mathbf{a}$ по обучающей совокупности и подставить в функцию потерь для контрольной совокупности: $\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x}) \to Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)$ – потери на мысленном контроле
Идея скрытой кросс-валидации: минимум математического ожидания потерь	$\iiint Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} =$ $\iiint Q\big(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\big)\Big\{\int \Phi(\tilde{\mathbf{y}}\mid\mathbf{x},\mathbf{a})\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})\,d\mathbf{a}\Big\}G^*(\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} \to \min(C)$

В реальности у наблюдателя имеется единственная обучающая выборка $(\mathbf{x},\mathbf{y})$.

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y} \mid \mathbf{x}, \mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,\ldots,y_N)\in\mathbb{R}^N$, $\tilde{\mathbf{y}}=(\tilde{y}_1,\ldots,\tilde{y}_N)\in\mathbb{R}^N$. Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая
Мысленная перекрестная проверка на генеральной совокупности	Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения C вычислить оценку $\mathbf{a}$ по обучающей совокупности и подставить в функцию потерь для контрольной совокупности: $\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x}) \to Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)$ – потери на мысленном контроле
Идея скрытой кросс-валидации: минимум математического ожидания потерь	$\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} =$ $\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)\left\{\int \Phi(\tilde{\mathbf{y}}\mid\mathbf{x},\mathbf{a})\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})\,d\mathbf{a}\right\}G^*(\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} \to \min(C)$

В реальности у наблюдателя имеется единственная обучающая выборка $(\mathbf{x},\mathbf{y})$.

Он может подставить в функцию потерь лишь оценку, вычисленную по той же выборке $Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right)$.

Принцип неявной кросс-валидации: Мысленный эксперимент наблюдателя

Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых компонент согласно $G^*(\mathbf{x})$	Затем, дважды применив условное распределение $\Phi(\mathbf{y} \mid \mathbf{x}, \mathbf{a})$, получила две реализации $\mathbf{y}=(y_1,\ldots,y_N)\in\mathbb{R}^N$, $\tilde{\mathbf{y}}=(\tilde{y}_1,\ldots,\tilde{y}_N)\in\mathbb{R}^N$. Две мысленные совокупности: $(\mathbf{x},\mathbf{y})$ – контрольная, $(\mathbf{x},\tilde{\mathbf{y}})$ – обучающая
Мысленная перекрестная проверка на генеральной совокупности	Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения C вычислить оценку $\mathbf{a}$ по обучающей совокупности и подставить в функцию потерь для контрольной совокупности: $\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x}) \to Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)$ – потери на мысленном контроле
Идея скрытой кросс-валидации: минимум математического ожидания потерь	$\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} =$ $\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)\left\{\int \Phi(\tilde{\mathbf{y}}\mid\mathbf{x},\mathbf{a})\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})\,d\mathbf{a}\right\}G^*(\mathbf{x})\,d\tilde{\mathbf{y}}\,d\mathbf{y}\,d\mathbf{x} \to \min(C)$

В реальности у наблюдателя имеется единственная обучающая выборка $(\mathbf{x},\mathbf{y})$.

Он может подставить в функцию потерь лишь оценку, вычисленную по той же выборке $Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right)$.

Насколько испортится критерий, подлежащий максимизации?
Каким должен быть ***штраф*** за использование оценки параметра, вычисленного ***по той же выборке***?

Штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке

Штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке

Еще раз идея скрытой кросс-валидации: минимум двойного математического ожидания	$\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x}) d\tilde{\mathbf{y}} d\mathbf{y} d\mathbf{x} =$ $\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) \left\{ \int \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Psi^*(\mathbf{a}) d\mathbf{a} \right\} G^*(\mathbf{x}) d\tilde{\mathbf{y}}\, d\mathbf{y}\, d\mathbf{x} \rightarrow$ $\rightarrow \min(C)$

Штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке

Еще раз идея скрытой кросс-валидации: минимум двойного математического ожидания	$\iiint Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})\, d\tilde{\mathbf{y}}\, d\mathbf{y}\, d\mathbf{x} =$ $\iiint Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)\left\{\int \Phi(\tilde{\mathbf{y}}\,|\,\mathbf{x},\mathbf{a})\Phi(\tilde{\mathbf{y}}\,|\,\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})\, d\mathbf{a}\right\} G^*(\mathbf{x})\, d\tilde{\mathbf{y}}\, d\mathbf{y}\, d\mathbf{x} \to$ $\to \min(C)$
Добавление нуля к подынтегральной функции	$Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) + \left\{Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) - Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right)\right\}$

Штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке

Еще раз идея скрытой кросс-валидации: минимум двойного математического ожидания	$\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x}) d\tilde{\mathbf{y}} d\mathbf{y} d\mathbf{x} =$ $\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) \left\{ \int \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Psi^*(\mathbf{a}) d\mathbf{a} \right\} G^*(\mathbf{x}) d\tilde{\mathbf{y}}\, d\mathbf{y}\, d\mathbf{x} \rightarrow$ $\rightarrow \min(C)$
Добавление нуля к подынтегральной функции	$Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) + \left\{ Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) - Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) \right\}$

Эквивалентная запись критерия минимизации по значению структурного параметра C :

$$C^* = \arg\min_C \left\{ \int\int \left[\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) \Phi(\mathbf{y} \mid \mathbf{x},\mathbf{a}) d\,\mathbf{y} + \right.\right.$$

$$\left.\left. \underbrace{\int\int \left(Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) - Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) \right) \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Phi(\mathbf{y} \mid \mathbf{x},\mathbf{a}) d\tilde{\mathbf{y}} d\mathbf{y}}_{\text{функционал от функции потерь } Q(\bullet,\mathbf{x},\mathbf{a}) \text{ и критерия обучения } \hat{\mathbf{a}}_C(\bullet,\mathbf{x})} \right] \Psi^*(\mathbf{a}) G^*(\mathbf{x}) d\mathbf{a} d\,\mathbf{x} \right\}$$

Штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке

Еще раз идея скрытой кросс-валидации: минимум двойного математического ожидания	$\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x}) d\tilde{\mathbf{y}} d\mathbf{y} d\mathbf{x} =$ $\int\int\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) \left\{ \int \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Psi^*(\mathbf{a}) d\mathbf{a} \right\} G^*(\mathbf{x}) d\tilde{\mathbf{y}}\, d\mathbf{y}\, d\mathbf{x} \to$ $\to \min(C)$
Добавление нуля к подынтегральной функции	$Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) + \left\{ Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) - Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) \right\}$

Эквивалентная запись критерия минимизации по значению структурного параметра C :

$$C^* = \arg\min_C \left\{ \int\int \left[\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) \Phi(\mathbf{y} \mid \mathbf{x},\mathbf{a}) d\,\mathbf{y} + \right.\right.$$

$$\left.\left. \underbrace{\int\int \left(Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) - Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) \right) \Phi(\tilde{\mathbf{y}} \mid \mathbf{x},\mathbf{a}) \Phi(\mathbf{y} \mid \mathbf{x},\mathbf{a}) d\tilde{\mathbf{y}} d\mathbf{y}}_{\text{функционал от функции потерь } Q(\bullet,\mathbf{x},\mathbf{a}) \text{ и критерия обучения } \hat{\mathbf{a}}_C(\bullet,\mathbf{x})} \right] \Psi^*(\mathbf{a}) G^*(\mathbf{x}) d\mathbf{a} d\,\mathbf{x} \right\}$$

Для многих типичных функций потерь $Q(\mathbf{y},\mathbf{x},\mathbf{a})$ и критериев обучения $\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})$, адекватных широкому классу практических задач, функционал во втором слагаемом не зависит от $\mathbf{a}$.	$\Delta(C,\mathbf{x})$

Штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке

Еще раз идея скрытой кросс-валидации: максимум двойного математического ожидания	$\iiint Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) F^*(\mathbf{y},\tilde{\mathbf{y}},\mathbf{x})\, d\tilde{\mathbf{y}}\, d\mathbf{y}\, d\mathbf{x} =$ $\iiint Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right)\left\{\int \Phi(\tilde{\mathbf{y}}\,|\,\mathbf{x},\mathbf{a})\Phi(\tilde{\mathbf{y}}\,|\,\mathbf{x},\mathbf{a})\Psi^*(\mathbf{a})\, d\mathbf{a}\right\} G^*(\mathbf{x})\, d\tilde{\mathbf{y}}\, d\mathbf{y}\, d\mathbf{x} \to$ $\to \min(C)$
Добавление нуля к подынтегральной функции	$Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\tilde{\mathbf{y}},\mathbf{x})\right) + \left\{Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right) - Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right)\right\}$

Эквивалентная запись критерия максимизации по значению структурного параметра C:

$$C^* = \arg\min_C \left\{ \iint \left[\int Q\left(\mathbf{y},\mathbf{x},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})\right)\Phi(\mathbf{y}\,|\,\mathbf{x},\mathbf{a})\, d\mathbf{y} + \right.\right.$$

$$\left.\left. + \Delta\left(C,\mathbf{x}\right) \right] \Psi^*(\mathbf{a}) G^*(\mathbf{x})\, d\mathbf{a}\, d\mathbf{x} \right\}$$

Для многих типичных функций потерь $Q(\mathbf{y},\mathbf{x},\mathbf{a})$ и критериев обучения $\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{x})$, адекватных широкому классу практических задач, функционал во втором слагаемом не зависит от $\mathbf{a}$.	$\Delta\left(C,\mathbf{x}\right)$

Вторая эвристика наблюдателя

Вторая эвристика наблюдателя

Теорема 1. Критерий скрытой кросс-валидации имеет вид:	$$C^* = \arg\min_C \left\{ \int\int \left[Q\left(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{x})\right) + \Delta\left(C, \mathbf{x}\right) \right] F^*(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} \right\}$$

Вторая эвристика наблюдателя

Теорема 1. Критерий скрытой кросс-валидации имеет вид:	$C^* = \arg\min_C \left\{ \int\int \left[Q\left(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{x})\right) + \Delta\left(C, \mathbf{x}\right) \right] F^*(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} \right\}$

Однако в таком виде критерий по-прежнему неприменим, так как распределение генеральной совокупности $F^*(\mathbf{x}, \mathbf{y})$ неизвестно.

Идея заключается в замене математического ожидания его несмещенной оценкой по единственной доступной наблюдателю выборке $\mathbf{y} \in \mathbb{R}^N$:

$$\hat{C}(\mathbf{y}, \mathbf{x}) = \arg\min_C \underbrace{\left\{ \underbrace{Q\left(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{x})\right)}_{\text{эмпирический риск}} + \Delta\left(C, \mathbf{x}\right) \right\}}_{\text{структурный риск}}$$

Вторая эвристика наблюдателя

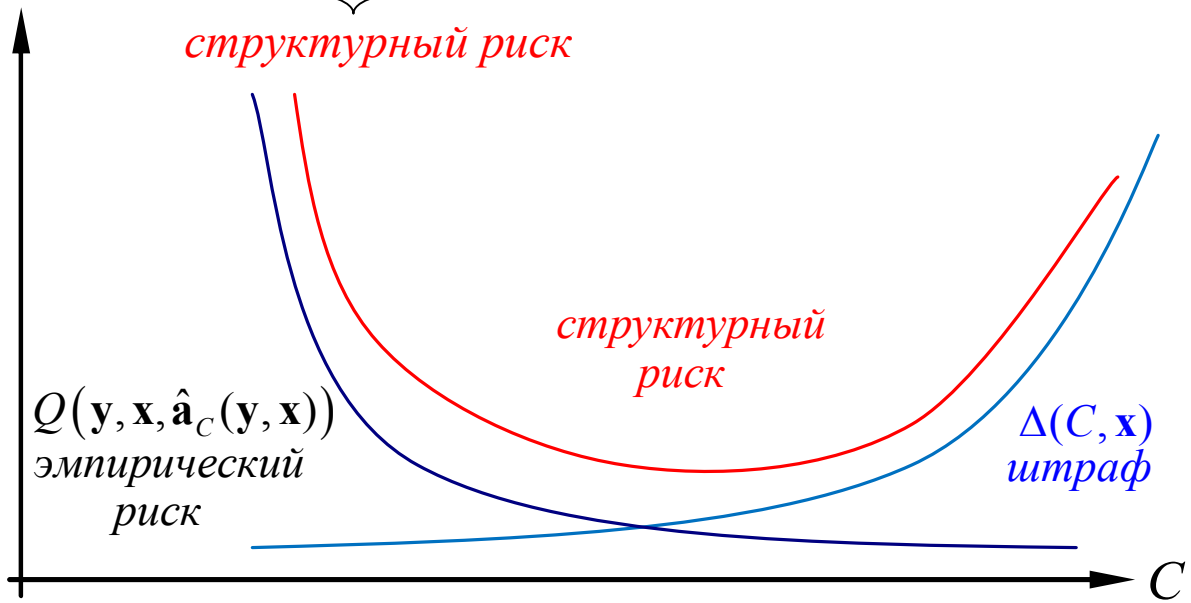
Теорема 1. Критерий скрытой кросс-валидации имеет вид:	$C^* = \arg\min_C \left\{ \int\int \left[Q\left(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{x})\right) + \Delta\left(C, \mathbf{x}\right) \right] F^*(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} \right\}$

Однако в таком виде критерий по-прежнему неприменим, так как распределение генеральной совокупности $F^*(\mathbf{x}, \mathbf{y})$ неизвестно.

Идея заключается в замене математического ожидания его несмещенной оценкой по единственной доступной наблюдателю выборке $\mathbf{y} \in \mathbb{R}^N$:

$$\hat{C}(\mathbf{y}, \mathbf{x}) = \arg\min_C \underbrace{\left\{ \underbrace{Q\left(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{x})\right)}_{\text{эмпирический риск}} + \Delta\left(C, \mathbf{x}\right) \right\}}_{\text{структурный риск}}$$

Графическое отображение зависимости суммарного структурного риска от параметра регуляризации в предположении, что свобода выбора решающего правила увеличивается при увеличении параметра C.



Оценивание числовой зависимости. Линейно-квадратичная модель

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: действительный вектор $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathbb{R}^n$.

Скрытая характеристика: действительное число $y(\omega) \in \mathbb{R}$.

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: действительный вектор $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathbb{R}^n$.

Скрытая характеристика: действительное число $y(\omega) \in \mathbb{R}$.

Центрированная и нормированная обучающая совокупность

$(\mathbf{X}, \mathbf{y}) = \left\{(\boldsymbol{x}_j, y_j),\ j = 1, \ldots, N\right\}$, $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N)$ – матрица $(n \times N)$, $\mathbf{y} \in \mathbb{R}^N$.

$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0,\ \frac{1}{N}\sum_{j=1}^{N} x_{ij}^2 = 1$ для всех $i = 1, \ldots, n$.

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: действительный вектор $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathbb{R}^n$.

Скрытая характеристика: действительное число $y(\omega) \in \mathbb{R}$

Центрированная и нормированная обучающая совокупность

$(\mathbf{X}, \mathbf{y}) = \left\{(\boldsymbol{x}_j, y_j),\ j = 1,\ldots,N\right\}$, $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N)$ – матрица $(n \times N)$, $\mathbf{y} \in \mathbb{R}^N$.

$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0,\ \frac{1}{N}\sum_{j=1}^{N} x_{ij}^2 = 1$ для всех $i = 1,\ldots,n$.

Искомое линейное решающее правило: $\hat{y}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$.

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: действительный вектор $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathbb{R}^n$.

Скрытая характеристика: действительное число $y(\omega) \in \mathbb{R}$

Центрированная и нормированная обучающая совокупность

$(\mathbf{X}, \mathbf{y}) = \left\{(\boldsymbol{x}_j, y_j),\ j = 1, \ldots, N\right\}$, $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N)$ – матрица $(n \times N)$, $\mathbf{y} \in \mathbb{R}^N$.

$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0,\ \frac{1}{N}\sum_{j=1}^{N} x_{ij}^2 = 1$ для всех $i = 1, \ldots, n$.

Искомое линейное решающее правило: $\hat{y}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$.

Квадратичная функция потерь: $Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) = \sum_{j=1}^{N} (y_j - \mathbf{a}^T\mathbf{x}_j)^2 = (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T(\mathbf{y} - \mathbf{X}^T\mathbf{a})$

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: действительный вектор $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathbb{R}^n$.

Скрытая характеристика: действительное число $y(\omega) \in \mathbb{R}$

Центрированная и нормированная обучающая совокупность

$(\mathbf{X}, \mathbf{y}) = \left\{(\boldsymbol{x}_j, y_j),\ j = 1,\ldots,N\right\}$, $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N)$ – матрица $(n \times N)$, $\mathbf{y} \in \mathbb{R}^N$.

$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0,\ \frac{1}{N}\sum_{j=1}^{N} x_{ij}^2 = 1$ для всех $i = 1,\ldots,n$.

Искомое линейное решающее правило: $\hat{y}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$.

Квадратичная функция потерь: $Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) = \sum_{j=1}^{N} (y_j - \mathbf{a}^T\mathbf{x}_j)^2 = (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T(\mathbf{y} - \mathbf{X}^T\mathbf{a})$

Напомним общий критерий обучения: $\hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{X}) = \arg\min_{\mathbf{a}} \left\{ \underbrace{V(\mathbf{a}, C)}_{\substack{\textit{регуляризующая} \\ \textit{функция}}} + Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) \right\}$

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: действительный вектор $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathbb{R}^n$.

Скрытая характеристика: действительное число $y(\omega) \in \mathbb{R}$

Центрированная и нормированная обучающая совокупность

$(\mathbf{X}, \mathbf{y}) = \left\{(\boldsymbol{x}_j, y_j),\ j = 1,\ldots,N\right\}$, $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N)$ – матрица $(n \times N)$, $\mathbf{y} \in \mathbb{R}^N$.

$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0,\ \frac{1}{N}\sum_{j=1}^{N} x_{ij}^2 = 1$ для всех $i = 1,\ldots,n$.

Искомое линейное решающее правило: $\hat{y}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$.

Квадратичная функция потерь: $Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) = \sum_{j=1}^{N} (y_j - \mathbf{a}^T\mathbf{x}_j)^2 = (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T(\mathbf{y} - \mathbf{X}^T\mathbf{a})$

Напомним общий критерий обучения: $\hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{X}) = \arg\min_{\mathbf{a}} \left\{ \underbrace{V(\mathbf{a}, C)}_{\substack{\textit{регуляризующая} \\ \textit{функция}}} + Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) \right\}$

Квадратичная регуляризация: $V(\mathbf{a}, C) = \mathbf{a}^T\mathbf{B}_C\mathbf{a}$,

$\mathbf{B}_C$ – квадратная матрица $(n \times n)$, определяемая конкретной прикладной задачей.

Оценивание числовой зависимости. Линейно-квадратичная модель

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: действительный вектор $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathbb{R}^n$.

Скрытая характеристика: действительное число $y(\omega) \in \mathbb{R}$

Центрированная и нормированная обучающая совокупность

$(\mathbf{X}, \mathbf{y}) = \left\{(\boldsymbol{x}_j, y_j),\ j = 1,\ldots,N\right\}$, $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N)$ – матрица $(n \times N)$, $\mathbf{y} \in \mathbb{R}^N$.

$\frac{1}{N}\sum_{j=1}^{N} x_{ij} = 0,\ \frac{1}{N}\sum_{j=1}^{N} x_{ij}^2 = 1$ для всех $i = 1,\ldots,n$.

Искомое линейное решающее правило: $\hat{y}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$.

Квадратичная функция потерь: $Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) = \sum_{j=1}^{N} (y_j - \mathbf{a}^T\mathbf{x}_j)^2 = (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T(\mathbf{y} - \mathbf{X}^T\mathbf{a})$

Напомним общий критерий обучения: $\hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{X}) = \arg\min_{\mathbf{a}} \left\{ \underbrace{V(\mathbf{a}, C)}_{\substack{\textit{регуляризующая} \\ \textit{функция}}} + Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) \right\}$

Квадратичная регуляризация: $V(\mathbf{a}, C) = \mathbf{a}^T\mathbf{B}_C\mathbf{a}$,

$\mathbf{B}_C$ – квадратная матрица $(n \times n)$, определяемая конкретной прикладной задачей.

Итак, квадратичная задача обучения: $\hat{\mathbf{a}}_C(\mathbf{y}, \mathbf{X}) = \arg\min_{\mathbf{a}} \left\{ \mathbf{a}^T\mathbf{B}_C\mathbf{a} + (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T(\mathbf{y} - \mathbf{X}^T\mathbf{a}) \right\}$

Очевидно, что задача обучения сводится к решению системы n линейных уравнений.

Оценивание числовой зависимости. Линейно-квадратичная модель

Итак, квадратичная задача обучения для заданного значения параметра регуляризации C:

$$\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X}) = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \left\{ \mathbf{a}^T \mathbf{B}_C \mathbf{a} + (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T (\mathbf{y} - \mathbf{X}^T\mathbf{a}) \right\}$$

Напомним общий вид критерия неявной кросс-валидации:

$$\hat{C}(\mathbf{y},\mathbf{X}) = \arg\min_{C} \left\{ Q\left(\mathbf{y},\mathbf{X},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X})\right) + \Delta\left(C,\mathbf{X}\right) \right\}.$$

Оценивание числовой зависимости. Линейно-квадратичная модель

Итак, квадратичная задача обучения для заданного значения параметра регуляризации C:

$$\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X}) = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \left\{ \mathbf{a}^T \mathbf{B}_C \mathbf{a} + (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T (\mathbf{y} - \mathbf{X}^T\mathbf{a}) \right\}$$

Напомним общий вид критерия неявной кросс-валидации:

$$\hat{C}(\mathbf{y},\mathbf{X}) = \arg\min_{C} \left\{ Q\left(\mathbf{y},\mathbf{X},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X})\right) + \Delta\left(C,\mathbf{X}\right) \right\}.$$

Напомним также первую эвристику наблюдателя в нашей теории:

Принцип незлонамеренности природы. Наблюдатель правильно «угадал» функцию потерь и класс решающих правил. Природа чаще генерирует пары с низким значением штрафа.	$\Phi(\mathbf{y} \mid \mathbf{X},\mathbf{a}) \propto \exp\left\{-\underbrace{Q(\mathbf{X},\mathbf{y},\mathbf{a})}\right\}$ *принятая наблюдателем функция потерь*

Оценивание числовой зависимости. Линейно-квадратичная модель

Итак, квадратичная задача обучения для заданного значения параметра регуляризации C:

$$\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X}) = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \left\{ \mathbf{a}^T \mathbf{B}_C \mathbf{a} + (\mathbf{y} - \mathbf{X}^T\mathbf{a})^T (\mathbf{y} - \mathbf{X}^T\mathbf{a}) \right\}$$

Напомним общий вид критерия неявной кросс-валидации:

$$\hat{C}(\mathbf{y},\mathbf{X}) = \arg\min_{C} \left\{ Q\left(\mathbf{y},\mathbf{X},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X})\right) + \Delta\left(C,\mathbf{X}\right) \right\}$$

Напомним также первую эвристику наблюдателя в нашей теории:

Принцип незлонамеренности природы. Наблюдатель правильно «угадал» функцию потерь и класс решающих правил. Природа чаще генерирует пары с низким значением штрафа.	$\Phi(\mathbf{y} \mid \mathbf{X},\mathbf{a}) \propto \exp\left\{-\underbrace{Q(\mathbf{X},\mathbf{y},\mathbf{a})}\right\}$ *принятая наблюдателем функция потерь*

Теорема 2. В методе неявной кросс-валидации для линейно-квадратичной модели штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке, определяется выражением $\Delta(C,\mathbf{X}) = Tr\left[\mathbf{X}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T + \mathbf{B}_C \right)^{-1} \right]$.

Оценивание числовой зависимости. Линейно-квадратичная модель

Итак, квадратичная задача обучения для заданного значения параметра регуляризации C:

$$\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X}) = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \left\{ \mathbf{a}^T\mathbf{B}_C\mathbf{a} + (\mathbf{y}-\mathbf{X}^T\mathbf{a})^T(\mathbf{y}-\mathbf{X}^T\mathbf{a}) \right\}$$

Напомним общий вид критерия неявной кросс-валидации:

$$\hat{C}(\mathbf{y},\mathbf{X}) = \arg\min_{C} \left\{ Q\left(\mathbf{y},\mathbf{X},\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X})\right) + \Delta\left(C,\mathbf{X}\right) \right\}$$

Напомним также первую эвристику наблюдателя в нашей теории:

Принцип незлонамеренности природы. Наблюдатель правильно «угадал» функцию потерь и класс решающих правил. Природа чаще генерирует пары с низким значением штрафа.	$\Phi(\mathbf{y}\mid\mathbf{X},\mathbf{a}) \propto \exp\left\{-\underbrace{Q(\mathbf{X},\mathbf{y},\mathbf{a})}\right\}$ *принятая наблюдателем функция потерь*

Теорема 2. В методе неявной кросс-валидации для линейно-квадратичной модели штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке, определяется выражением $\Delta(C,\mathbf{X}) = Tr\left[\mathbf{X}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T+\mathbf{B}_C\right)^{-1}\right]$.

Критерий неявной кросс-валидации для линейно-квадратичной модели:

$\hat{C}(\mathbf{y},\mathbf{X}) = \arg\min_{C} \left\{ \left(\mathbf{y}-\mathbf{X}^T\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X})\right)^T\left(\mathbf{y}-\mathbf{X}^T\hat{\mathbf{a}}_C(\mathbf{y},\mathbf{X})\right) + Tr\left[\mathbf{X}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T+\mathbf{B}_C\right)^{-1}\right] \right\}$	$0 \le Tr\left[\mathbf{X}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T+\mathbf{B}_{\hat{C}(\mathbf{y},\mathbf{X})}\right)^{-1}\right] \le n$ эффективная дробная размерность оцененного вектора коэффициентов

Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Подмножество «полезных» признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$.

Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Подмножество «полезных» признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$.

Скрытая характеристика: индекс класса $y(\omega) \in \{-1, 1\}$.

Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Подмножество «полезных» признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$.

Скрытая характеристика: Индекс класса $y(\omega) \in \{-1, 1\}$

Параметрическое семейство разделяющих гиперплоскостей

$$\boldsymbol{a}^T \boldsymbol{x} + b \begin{cases} > 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = 1 \\ < 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = -1 \end{cases}$$

$$\boldsymbol{x} = (x_i, i \in \hat{\mathbb{I}}) \in \mathbb{R}^{\hat{n}}, \boldsymbol{a} = (a_i, i \in \hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}) \in \mathbb{R}^{\hat{n}}, b \in \mathbb{R}, |\hat{\mathbb{I}}| = \hat{n}$$

Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Подмножество «полезных» признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$.

Скрытая характеристика: индекс класса $y(\omega) \in \{-1, 1\}$

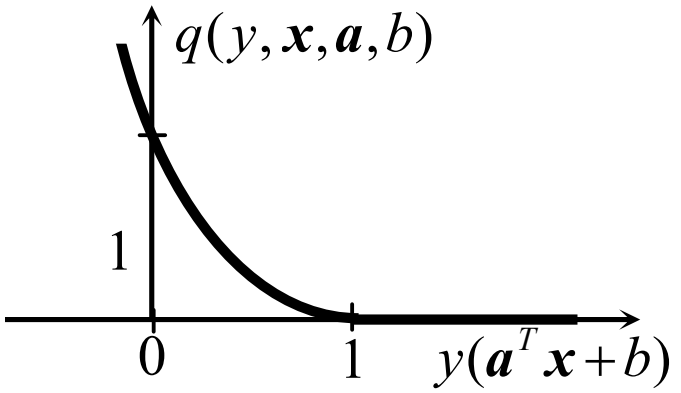
Параметрическое семейство разделяющих гиперплоскостей

$$\boldsymbol{a}^T \boldsymbol{x} + b \begin{cases} > 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = 1 \\ < 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = -1 \end{cases}$$

$$\boldsymbol{x} = (x_i, i \in \hat{\mathbb{I}}) \in \mathbb{R}^{\hat{n}}, \boldsymbol{a} = (a_i, i \in \hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}) \in \mathbb{R}^{\hat{n}}, b \in \mathbb{R}, |\hat{\mathbb{I}}| = \hat{n}$$

Функция потерь, применимая к любому объекту $(\boldsymbol{x}, y)$

$$q(y, \boldsymbol{x}, \boldsymbol{a}, b) = \begin{cases} 0, & \text{if } y(\boldsymbol{a}^T \boldsymbol{x} + b) > 1, \\ \left(1 - y(\boldsymbol{a}^T \boldsymbol{x} + b)\right)^2, & \text{if } y(\boldsymbol{a}^T \boldsymbol{x} + b) < 1, \end{cases}$$



Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Подмножество «полезных» признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$.

Скрытая характеристика: Индекс класса $y(\omega) \in \{-1, 1\}$

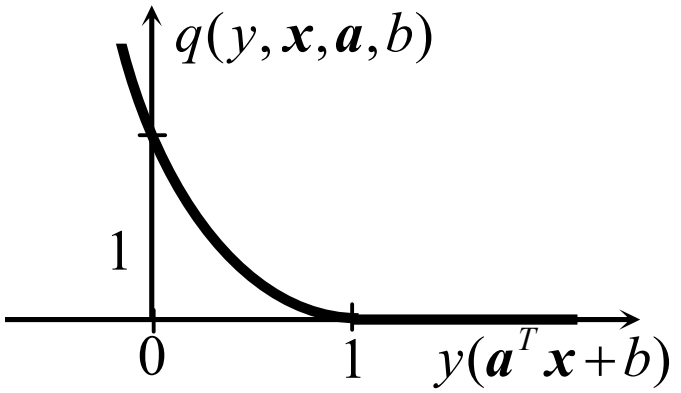
Параметрическое семейство разделяющих гиперплоскостей

$$\boldsymbol{a}^T\boldsymbol{x}+b \begin{cases} > 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = 1 \\ < 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = -1 \end{cases}$$

$$\boldsymbol{x} = (x_i, i \in \hat{\mathbb{I}}) \in \mathbb{R}^{\hat{n}}, \boldsymbol{a} = (a_i, i \in \hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}) \in \mathbb{R}^{\hat{n}}, b \in \mathbb{R}, |\hat{\mathbb{I}}| = \hat{n}$$

Функция потерь, применимая к любому объекту $(\boldsymbol{x}, y)$

$$q(y, \boldsymbol{x}, \boldsymbol{a}, b) = \begin{cases} 0, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b) > 1, \\ \left(1 - y(\boldsymbol{a}^T\boldsymbol{x}+b)\right)^2, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b) < 1, \end{cases}$$



Критерий обучения (квадратичная версия SVM):

$$\begin{pmatrix} \hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) \\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) \end{pmatrix} = \arg\min_{\boldsymbol{a},b} \left[\boldsymbol{a}^T\boldsymbol{a} + C \underbrace{\sum_{j:\, y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1} \left(y_j - (\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2}_{\text{эмпирический риск}} \right]$$

Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Подмножество «полезных» признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$.

Скрытая характеристика: Индекс класса $y(\omega) \in \{-1, 1\}$

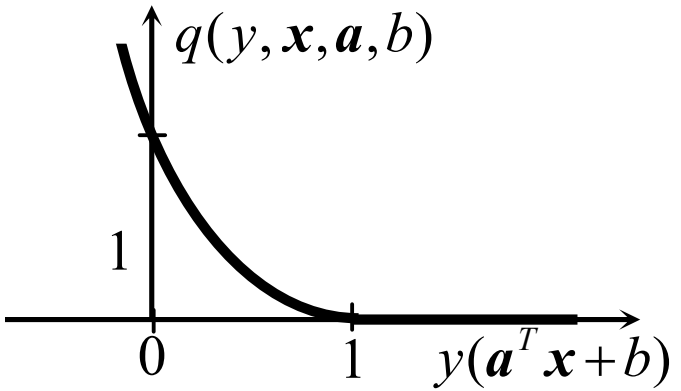
Параметрическое семейство разделяющих гиперплоскостей

$$\boldsymbol{a}^T\boldsymbol{x}+b \begin{cases} > 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = 1 \\ < 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = -1 \end{cases}$$

$$\boldsymbol{x} = (x_i, i \in \hat{\mathbb{I}}) \in \mathbb{R}^{\hat{n}}, \boldsymbol{a} = (a_i, i \in \hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}) \in \mathbb{R}^{\hat{n}}, b \in \mathbb{R}, |\hat{\mathbb{I}}| = \hat{n}$$

Функция потерь, применимая к любому объекту $(\boldsymbol{x}, y)$

$$q(y, \boldsymbol{x}, \boldsymbol{a}, b) = \begin{cases} 0, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b) > 1, \\ \left(1 - y(\boldsymbol{a}^T\boldsymbol{x}+b)\right)^2, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b) < 1, \end{cases}$$



Критерий обучения (квадратичная версия SVM):

$$\begin{pmatrix} \hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) \\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) \end{pmatrix} = \arg\min_{\boldsymbol{a},b} \left[\boldsymbol{a}^T\boldsymbol{a} + C \underbrace{\sum_{j:\, y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1} \left(y_j - (\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2}_{\text{эмпирический риск}} \right]$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$, $\boldsymbol{a}, \boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Обучение распознаванию образов. Метод опорных векторов

Некоторое множество реально существующих объектов $\omega \in \Omega$.

Наблюдаемая характеристика: Исходный вектор признаков $\boldsymbol{x}(\omega) = \left(x_i(\omega), i = 1,...,n\right)$.

Подмножество «полезных» признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$.

Скрытая характеристика: Индекс класса $y(\omega) \in \{-1, 1\}$

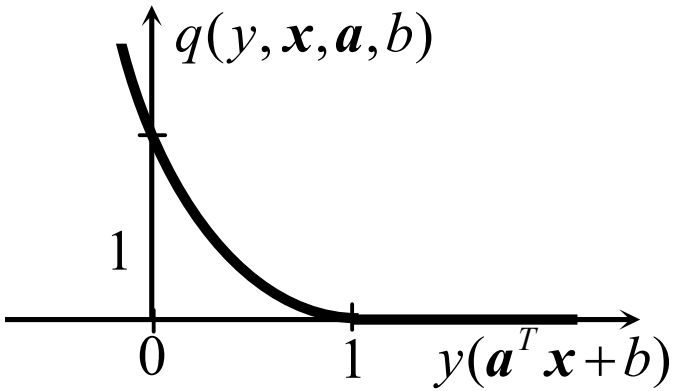
Параметрическое семейство разделяющих гиперплоскостей

$$\boldsymbol{a}^T\boldsymbol{x}+b \begin{cases} > 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = 1 \\ < 0 \Rightarrow \hat{y}(\boldsymbol{x} \mid \boldsymbol{a}, b) = -1 \end{cases}$$

$$\boldsymbol{x} = (x_i, i \in \hat{\mathbb{I}}) \in \mathbb{R}^{\hat{n}}, \boldsymbol{a} = (a_i, i \in \hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}) \in \mathbb{R}^{\hat{n}}, b \in \mathbb{R}, |\hat{\mathbb{I}}| = \hat{n}$$

Функция потерь, применимая к любому объекту $(\boldsymbol{x}, y)$

$$q(y, \boldsymbol{x}, \boldsymbol{a}, b) = \begin{cases} 0, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b) > 1, \\ \left(1 - y(\boldsymbol{a}^T\boldsymbol{x}+b)\right)^2, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b) < 1, \end{cases}$$



Критерий обучения (квадратичная версия SVM):

$$\begin{pmatrix} \hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) \\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) \end{pmatrix} = \arg\min_{\boldsymbol{a},b} \left[\boldsymbol{a}^T\boldsymbol{a} + C \underbrace{\sum_{j:\, y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1} \left(y_j - (\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2}_{\text{эмпирический риск}} \right]$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$, $\boldsymbol{a}, \boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Как выбрать структурные параметры по единственной обучающей совокупности?

Обучение распознаванию образов. Метод опорных векторов

Еще раз критерий обучения (квадратичная версия SVM):

$$\begin{pmatrix} \hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \end{pmatrix} = \arg\min_{\boldsymbol{a},b} \left[\boldsymbol{a}^T\boldsymbol{a} + C \underbrace{\sum_{j:\, y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1} \left(y_j - (\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2}_{\textit{эмпирический риск}} \right]$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,\ldots,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$, $\boldsymbol{a},\boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C.

Обучение распознаванию образов. Метод опорных векторов

Еще раз критерий обучения (квадратичная версия SVM):

$$\begin{pmatrix} \hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \end{pmatrix} = \arg\min_{\boldsymbol{a},b} \left[\boldsymbol{a}^T\boldsymbol{a} + C \underbrace{\sum_{j:\, y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1} \left(y_j - (\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2}_{\textit{эмпирический риск}} \right]$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,\ldots,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \le n$, $\boldsymbol{a},\boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Объединим параметры дискриминантной гиперплоскости: $\mathbf{a} = (\boldsymbol{a}, b) \in \mathbb{R}^{\hat{n}+1}$, $\mathbf{x} = (\boldsymbol{x}, 1) \in \mathbb{R}^{\hat{n}+1}$

Критерий обучения:

$$\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a}, \xi_j, j \in \mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \ge 1 - \xi_j,\ \xi_j \ge 0,\ j \in \mathbb{J} = \{1,\ldots,N\}, \end{cases} \qquad \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}.$$

Обучение распознаванию образов. Метод опорных векторов

Еще раз критерий обучения (квадратичная версия SVM):

$$\begin{pmatrix} \hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \end{pmatrix} = \arg\min_{\boldsymbol{a},b} \left[\boldsymbol{a}^T\boldsymbol{a} + C \underbrace{\sum_{j:\, y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1} \left(y_j - (\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2}_{\textit{эмпирический риск}} \right]$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,\ldots,n\},\ \hat{n} = |\hat{\mathbb{I}}| \le n,$ $\boldsymbol{a},\boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Объединим параметры дискриминантной гиперплоскости: $\mathbf{a} = (\boldsymbol{a}, b) \in \mathbb{R}^{\hat{n}+1}$, $\mathbf{x} = (\boldsymbol{x}, 1) \in \mathbb{R}^{\hat{n}+1}$

Критерий обучения:

$$\begin{cases} \mathbf{a}^T\mathbf{B}_{\hat{\mathbb{I}},C}\mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a}, \xi_j, j\in\mathbb{J}), \\ y_j\mathbf{x}_j^T\mathbf{a} \ge 1-\xi_j,\ \xi_j \ge 0,\ j\in\mathbb{J} = \{1,\ldots,N\}, \end{cases} \qquad \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$$

Решение:

активные ограничения

1) подмножество опорных объектов

$$\begin{cases} \hat{\mathbb{J}}_{\hat{\mathbb{I}},C} = \left\{ j\in\mathbb{J} : y_j\mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) = 1-\hat{\xi}_j,\ \hat{\xi}_j > 0 \right\}, \\ \hat{\xi}_j^2 = (y_j - \mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2. \end{cases}$$

2) параметры гиперплоскости $\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) = \begin{pmatrix} \hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) \end{pmatrix} = (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}.$

Здесь:

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} = \begin{pmatrix} \boldsymbol{x}_{j_1} \cdots \boldsymbol{x}_{j_{\hat{N}_{\hat{\mathbb{I}},C}}} \\ 1 \cdots 1 \end{pmatrix} \left((\hat{n}+1)\times\hat{N}_{\hat{\mathbb{I}},C}\right), \quad \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} = (y_{j_1}\cdots y_{j_{\hat{N}_{\hat{\mathbb{I}},C}}})^T, \quad \hat{N}_{\hat{\mathbb{I}},C} = |\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|$$

Обучение распознаванию образов. Метод опорных векторов

Критерий обучения (квадратичная версия SVM):

$$\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j \in \mathbb{J}} \xi_j^2 \to min(\mathbf{a}, \xi_j, j \in \mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1 - \xi_j, \xi_j \geq 0, j \in \mathbb{J} = \{1, \ldots, N\}, \end{cases} \quad \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1, \ldots, n\}$, $\hat{n} = |\hat{\mathbb{I}}| \leq n$, $\boldsymbol{a}, \boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Подмножество опорных объектов

$$\begin{cases} \hat{\mathbb{J}}_{\hat{\mathbb{I}},C} = \left\{ j \in \mathbb{J} : y_j \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) = 1 - \hat{\xi}_j, \hat{\xi}_j > 0 \right\}, \\ \hat{\xi}_j^2 = (y_j - \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2. \end{cases}$$

Дополнительные обозначения:

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} = \begin{pmatrix} \boldsymbol{x}_{j_1} \cdots \boldsymbol{x}_{j_{\hat{N}_{\hat{\mathbb{I}},C}}} \\ 1 \cdots 1 \end{pmatrix} \Big((\hat{n}+1) \times \hat{N}_{\hat{\mathbb{I}},C} \Big), \quad \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} = (y_{j_1} \cdots y_{j_{\hat{N}_{\hat{\mathbb{I}},C}}})^T, \quad \hat{N}_{\hat{\mathbb{I}},C} = |\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|.$$

Оптимальная гиперплоскость полностью определяется подмножеством опорных объектов

$$\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) = \operatorname{argmin} \Big(\mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j \in \hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} (y_j - \mathbf{x}_j^T \mathbf{a})^2 \Big) =$$

$$\operatorname{argmin} \left(\mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \left\| \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} - \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T \mathbf{a} \right\|^2 \right) = (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}$$

Обучение распознаванию образов. Метод опорных векторов

Критерий обучения (квадратичная версия SVM):

$$\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a}, \xi_j, j\in\mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1-\xi_j, \xi_j \geq 0, j\in\mathbb{J}=\{1,...,N\}, \end{cases} \quad \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,...,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \leq n$, $\boldsymbol{a}, \boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Подмножество опорных объектов

$$\begin{cases} \hat{\mathbb{J}}_{\hat{\mathbb{I}},C} = \left\{ j\in\mathbb{J} : y_j \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) = 1-\hat{\xi}_j, \hat{\xi}_j > 0 \right\}, \\ \hat{\xi}_j^2 = (y_j - \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2. \end{cases}$$

Дополнительные обозначения:

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} = \begin{pmatrix} \boldsymbol{x}_{j_1} \cdots \boldsymbol{x}_{j_{\hat{N}_{\hat{\mathbb{I}},C}}} \\ 1 \cdots 1 \end{pmatrix} \Big(\hat{n}+1 \Big) \times \hat{N}_{\hat{\mathbb{I}},C} \Big), \quad \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} = (y_{j_1} \cdots y_{j_{\hat{N}_{\hat{\mathbb{I}},C}}})^T, \quad \hat{N}_{\hat{\mathbb{I}},C} = |\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|.$$

Оптимальная гиперплоскость полностью определяется подмножеством опорных объектов

$$\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) = \operatorname{argmin}\Big(\mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} (y_j - \mathbf{x}_j^T \mathbf{a})^2 \Big) =$$

$$\operatorname{argmin}\left(\mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \left\| \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} - \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T \mathbf{a} \right\|^2 \right) = (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}$$

Для сравнения:
Критерий обучения в линейно-квадратичной модели

$$\hat{\mathbf{a}}_C(\boldsymbol{X},\boldsymbol{y}) = \arg\min\left(\mathbf{a}^T \mathbf{B}_C \mathbf{a} + \left\| \boldsymbol{y} - \boldsymbol{X}^T \mathbf{a} \right\|^2 \right) = (\boldsymbol{X}\boldsymbol{X}^T + \mathbf{B}_C)^{-1} \boldsymbol{X}\boldsymbol{y}$$

Обучение распознаванию образов. Метод опорных векторов

Критерий обучения (квадратичная версия SVM):

$$\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a}, \xi_j, j\in\mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1-\xi_j, \xi_j \geq 0, j\in\mathbb{J}=\{1,\ldots,N\}, \end{cases} \quad \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1,\ldots,n\}$, $\hat{n} = |\hat{\mathbb{I}}| \leq n$, $\boldsymbol{a}, \boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Подмножество опорных объектов

$$\begin{cases} \hat{\mathbb{J}}_{\hat{\mathbb{I}},C} = \left\{ j\in\mathbb{J} : y_j \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) = 1-\hat{\xi}_j, \hat{\xi}_j > 0 \right\}, \\ \hat{\xi}_j^2 = (y_j - \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2. \end{cases}$$

Дополнительные обозначения:

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} = \begin{pmatrix} \boldsymbol{x}_{j_1} \cdots \boldsymbol{x}_{j_{\hat{N}_{\hat{\mathbb{I}},C}}} \\ 1 \cdots 1 \end{pmatrix} \left((\hat{n}+1)\times \hat{N}_{\hat{\mathbb{I}},C} \right), \quad \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} = (y_{j_1} \cdots y_{j_{\hat{N}_{\hat{\mathbb{I}},C}}})^T, \quad \hat{N}_{\hat{\mathbb{I}},C} = |\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|.$$

Оптимальная гиперплоскость полностью определяется подмножеством опорных объектов

$$\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y}) = \operatorname{argmin}\left(\mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} (y_j - \mathbf{x}_j^T \mathbf{a})^2 \right) =$$

$$\operatorname{argmin}\left(\mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \left\| \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} - \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T \mathbf{a} \right\|^2 \right) = (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}.$$

Центральная идея, сближающая модель SVM и линейную квадратичную модель:

Рассматривать подмножество опорных объектов $\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}$ как вторичный структурный параметр, полностью определяемый исходными структурными параметрами $(\hat{\mathbb{I}}, C)$

Критерий неявной кросс-валидации выбора структурных параметров в квадратичной модели SVM

Критерий обучения:

$$\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j \in \mathbb{J}} \xi_j^2 \to min(\mathbf{a}, \xi_j, j \in \mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1 - \xi_j, \xi_j \geq 0, j \in \mathbb{J} = \{1, \ldots, N\}, \end{cases} \quad \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$$

Структурные параметры:

1) подмножество признаков $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1, \ldots, n\}$, $\hat{n} = |\hat{\mathbb{I}}| \leq n$, $\boldsymbol{a}, \boldsymbol{x}_j \in \mathbb{R}^{\hat{n}}$

2) коэффициент баланса C

Подмножество опорных объектов, найденное в результате обучения

$$\begin{cases} \hat{\mathbb{J}}_{\hat{\mathbb{I}},C} = \left\{ j \in \mathbb{J} : y_j \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y}) = 1 - \hat{\xi}_j, \hat{\xi}_j > 0 \right\}, \\ \hat{\xi}_j^2 = (y_j - \mathbf{x}_j^T \hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2. \end{cases}$$

Дополнительные обозначения:

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} = \begin{pmatrix} \boldsymbol{x}_{j_1} \cdots \boldsymbol{x}_{j_{\hat{N}_{\hat{\mathbb{I}},C}}} \\ 1 \cdots 1 \end{pmatrix} \left(\hat{n}+1 \right) \times \hat{N}_{\hat{\mathbb{I}},C} \Big), \quad \hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C} = (y_{j_1} \cdots y_{j_{\hat{N}_{\hat{\mathbb{I}},C}}})^T, \quad \hat{N}_{\hat{\mathbb{I}},C} = |\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|.$$

Критерий неявной кросс-валидации для подмножества опорных объектов как вторичного структурного параметра, использующий аналогию с линейно-квадратичной моделью:

$$\begin{pmatrix} \hat{\mathbb{I}}(\boldsymbol{y}, \boldsymbol{X}) \\ \hat{C}(\boldsymbol{y}, \boldsymbol{X}) \end{pmatrix} = \underset{\hat{\mathbb{I}} \subseteq \mathbb{I},\, C}{\arg\min} \left\{ \sum_{j \in \hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}_{\hat{\mathbb{I}},C,j}^2 + \Delta(\hat{\mathbb{I}}, C, \boldsymbol{X}) \right\}, \quad \Delta(\hat{\mathbb{I}}, C, \boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right]$$

Критерий неявной кросс-валидации выбора структурных параметров в квадратичной модели SVM

Итак, критерий неявной кросс-валидации для подмножества опорных объектов как вторичного структурного параметра:

$$\begin{pmatrix} \hat{\mathbb{I}}(\boldsymbol{y},\boldsymbol{X}) \\ \hat{C}(\boldsymbol{y},\boldsymbol{X}) \end{pmatrix} = \arg\min_{\hat{\mathbb{I}}\subseteq\mathbb{I},\, C} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^{2}_{\hat{\mathbb{I}},C,j} + \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) \right\}, \quad \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^{T}_{\hat{\mathbb{I}},C} (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^{T}_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right]$$

Критерий неявной кросс-валидации выбора структурных параметров в квадратичной модели SVM

Итак, критерий неявной кросс-валидации для подмножества опорных объектов как вторичного структурного параметра:

$$\begin{pmatrix} \hat{\mathbb{I}}(\boldsymbol{y},\boldsymbol{X}) \\ \hat{C}(\boldsymbol{y},\boldsymbol{X}) \end{pmatrix} = \arg\min_{\mathbb{I}\subseteq\mathbb{I},\, C} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^2_{\hat{\mathbb{I}},C,j} + \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) \right\}, \quad \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right]$$

Напомним критерий обучения: $\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a},\xi_j, j\in\mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1-\xi_j,\ \xi_j \geq 0,\ j\in\mathbb{J}=\{1,\ldots,N\}, \end{cases} \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$

Критерий неявной кросс-валидации выбора структурных параметров в квадратичной модели SVM

Итак, критерий неявной кросс-валидации для подмножества опорных объектов как вторичного структурного параметра:

$$\begin{pmatrix} \hat{\mathbb{I}}(\boldsymbol{y},\boldsymbol{X}) \\ \hat{C}(\boldsymbol{y},\boldsymbol{X}) \end{pmatrix} = \arg\min_{\mathbb{I}\subseteq\mathbb{I},\, C} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^2_{\hat{\mathbb{I}},C,j} + \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) \right\}, \quad \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right]$$

Напомним критерий обучения: $\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a},\xi_j, j\in\mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1-\xi_j,\ \xi_j \geq 0,\ j\in\mathbb{J}=\{1,\ldots,N\}, \end{cases}$ $\underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$

Алгоритм решения задачи обучения расскажет Павел Левдик в следующем докладе.

Критерий неявной кросс-валидации выбора структурных параметров в квадратичной модели SVM

Итак, критерий неявной кросс-валидации для подмножества опорных объектов как вторичного структурного параметра:

$$\begin{pmatrix} \hat{\mathbb{I}}(\boldsymbol{y},\boldsymbol{X}) \\ \hat{C}(\boldsymbol{y},\boldsymbol{X}) \end{pmatrix} = \arg\min_{\mathbb{I}\subseteq\mathbb{I},\, C} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^2_{\hat{\mathbb{I}},C,j} + \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) \right\}, \quad \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right]$$

Напомним критерий обучения: $\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a},\xi_j, j\in\mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1-\xi_j,\ \xi_j \geq 0,\ j\in\mathbb{J}=\{1,\ldots,N\}, \end{cases}$ $\underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$

Алгоритм решения задачи обучения расскажет Павел Левдик в следующем докладе.

Опыт показывает, что значение параметра C следует принимать как можно бóльшим.

Пусть $C = const \to \infty$.

Теорема. $\lim\limits_{C\to\infty} \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = \lim\limits_{C\to\infty} Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right] = \min\left\{ \hat{n}, \hat{N}_{\hat{\mathbb{I}},C} \right\} + 1$

Критерий неявной кросс-валидации выбора структурных параметров в квадратичной модели SVM

Итак, критерий неявной кросс-валидации для подмножества опорных объектов как вторичного структурного параметра:

$$\begin{pmatrix} \hat{\mathbb{I}}(\boldsymbol{y},\boldsymbol{X}) \\ \hat{C}(\boldsymbol{y},\boldsymbol{X}) \end{pmatrix} = \arg\min_{\mathbb{I}\subseteq\mathbb{I},\, C} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^2_{\hat{\mathbb{I}},C,j} + \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) \right\}, \quad \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} \left(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C} \right)^{-1} \right]$$

Напомним критерий обучения: $\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a},\xi_j, j\in\mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1-\xi_j,\ \xi_j \geq 0,\ j\in\mathbb{J}=\{1,\ldots,N\}, \end{cases}$ $\underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$

Алгоритм решения задачи обучения расскажет Павел Левдик в следующем докладе.

Опыт показывает, что значение параметра C следует принимать как можно бóльшим.

Пусть $C = const \to \infty$.

Теорема. $\lim\limits_{C\to\infty} \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = \lim\limits_{C\to\infty} Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} \left(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C} \right)^{-1} \right] = \min\left\{ \hat{n},\, \hat{N}_{\hat{\mathbb{I}},C} \right\} + 1.$

Критерий неявной кросс-валидации для фиксированного большого значения C:

$$\hat{\mathbb{I}}_C(\boldsymbol{y},\boldsymbol{X}) = \arg\min_{\mathbb{I}\subseteq\mathbb{I}} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^2_{\hat{\mathbb{I}},C,j} + \left(\min\left\{ \hat{n},\, \hat{N}_{\hat{\mathbb{I}},\mathrm{C}} \right\} + 1 \right) \right\}.$$

Критерий неявной кросс-валидации выбора структурных параметров в квадратичной модели SVM

Итак, критерий неявной кросс-валидации для подмножества опорных объектов как вторичного структурного параметра:

$$\begin{pmatrix} \hat{\mathbb{I}}(\boldsymbol{y},\boldsymbol{X}) \\ \hat{C}(\boldsymbol{y},\boldsymbol{X}) \end{pmatrix} = \arg\min_{\hat{\mathbb{I}}\subseteq\mathbb{I},\, C} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^2_{\hat{\mathbb{I}},C,j} + \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) \right\}, \quad \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right]$$

Напомним критерий обучения: $\begin{cases} \mathbf{a}^T \mathbf{B}_{\hat{\mathbb{I}},C} \mathbf{a} + \sum_{j\in\mathbb{J}} \xi_j^2 \to min(\mathbf{a},\xi_j, j\in\mathbb{J}), \\ y_j \mathbf{x}_j^T \mathbf{a} \geq 1-\xi_j,\ \xi_j \geq 0,\ j\in\mathbb{J}=\{1,\ldots,N\}, \end{cases}$ $\underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}$

Алгоритм решения задачи обучения расскажет Павел Левдик в следующем докладе.

Опыт показывает, что значение параметра C следует принимать как можно бóльшим.

Пусть $C = const \to \infty$.

Теорема. $\lim\limits_{C\to\infty} \Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = \lim\limits_{C\to\infty} Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C} + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1} \right] = \min\left\{ \hat{n},\, \hat{N}_{\hat{\mathbb{I}},C} \right\} + 1$

Критерий неявной кросс-валидации для фиксированного большого значения C:

$$\hat{\mathbb{I}}_C(\boldsymbol{y},\boldsymbol{X}) = \arg\min_{\hat{\mathbb{I}}\subseteq\mathbb{I}} \left\{ \sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}^2_{\hat{\mathbb{I}},C,j} + \left(\min\left\{ \hat{n},\, \hat{N}_{\hat{\mathbb{I}},\mathrm{C}} \right\} + 1 \right) \right\}.$$

Конечно, огромное число 2^n всех подмножеств в множестве признаков $\hat{\mathbb{I}}\subseteq\mathbb{I}=\{1,\ldots,n\}$ делает невозможной «наивную» реализацию этого критерия.

В следующем докладе Павел Левдик расскажет, как обойтись вообще без перебора.

Спасибо за внимание!