Метод дифференциальной кросс-валидации для выбора уровня сложности обобщенных линейных моделей зависимостей

Ангуло Бриан Флориан
Московский физико-технический институт

Морозов Алексей Олегович
Московский физико-технический институт

Моттль Вадим Вячеславович
Вычислительный центр РАН

Обобщенная задача восстановления зависимости в линейном пространстве наблюдений

$\mathbf{x} \in \mathbb{R}^n$ – объекты реального мира, наблюдаемые через вещественные признаки

$y \in \mathbb{Y}$ – скрытое свойство объекта

$y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – неизвестная реально существующая зависимость

$\{(\mathbf{x}_j, y_j),\ j = 1,\ldots,N\}$ – множество прецедентов (обучающая совокупность)

$\hat{y} = \hat{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – требуется сконструировать решающее правило, применимое каждому $\mathbf{x} \in \mathbb{R}^n$, $\hat{y} \approx y$ (приближенно восстановить скрытую зависимость)

Если $\mathbb{Y} = \mathbb{R}$ оценивание регрессионной зависимости $y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$

Если $\mathbb{Y} = \{-1, 1\}$ двух-классовое распознавание образов $y = f(\mathbf{x}) : \mathbb{R}^n \to \{-1, 1\}$

Обобщенная линейная модель (Generalized Linear Model)

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$z(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T\mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$ обобщенная линейная модель зависимости

параметры модели: $\mathbf{a} \in \mathbb{R}^n$ – направляющий вектор, $b \in \mathbb{R}$ – сдвиг

Обобщенная задача восстановления зависимости в линейном пространстве наблюдений

$\mathbf{x} \in \mathbb{R}^n$ – объекты реального мира, наблюдаемые через вещественные признаки

$y \in \mathbb{Y}$ – скрытое свойство объекта

$y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – неизвестная реально существующая зависимость

$\{(\mathbf{x}_j, y_j),\ j = 1,\ldots,N\}$ – множество прецедентов (обучающая совокупность)

$\hat{y} = \hat{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – требуется сконструировать решающее правило, применимое каждому $\mathbf{x} \in \mathbb{R}^n$, $\hat{y} \approx y$ (приближенно восстановить скрытую зависимость)

Если $\mathbb{Y} = \mathbb{R}$ оценивание регрессионной зависимости $y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$

Если $\mathbb{Y} = \{-1, 1\}$ двух-классовое распознавание образов $y = f(\mathbf{x}) : \mathbb{R}^n \to \{-1, 1\}$

Обобщенная линейная модель (Generalized Linear Model)

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$z(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T\mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$ обобщенная линейная модель зависимости

$q(y, z) : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+$ функция связи (штрафа, потерь)

Обобщенная задача восстановления зависимости в линейном пространстве наблюдений

$\mathbf{x} \in \mathbb{R}^n$ – объекты реального мира, наблюдаемые через вещественные признаки

$y \in \mathbb{Y}$ – скрытое свойство объекта

$y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – неизвестная реально существующая зависимость

$\{(\mathbf{x}_j, y_j),\ j = 1,\ldots,N\}$ – множество прецедентов (обучающая совокупность)

$\hat{y} = \hat{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – требуется сконструировать решающее правило, применимое каждому $\mathbf{x} \in \mathbb{R}^n$, $\hat{y} \approx y$ (приближенно восстановить скрытую зависимость)

Если $\mathbb{Y} = \mathbb{R}$ оценивание регрессионной зависимости $y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$

Если $\mathbb{Y} = \{-1, 1\}$ двух-классовое распознавание образов $y = f(\mathbf{x}) : \mathbb{R}^n \to \{-1, 1\}$

Обобщенная линейная модель (Generalized Linear Model)

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$z(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T\mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$ обобщенная линейная модель зависимости

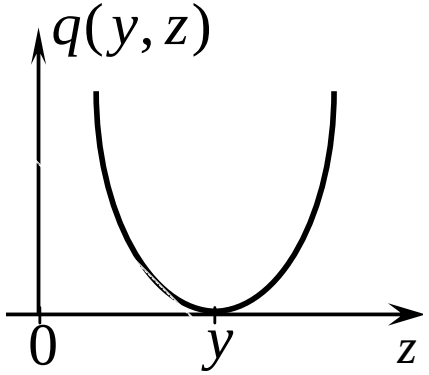
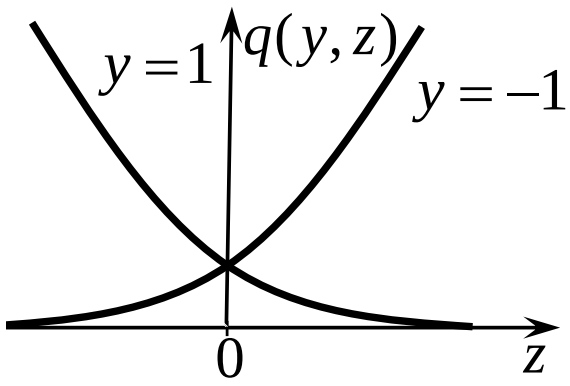
$q(y, z) : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+$ функция связи (штрафа, потерь)

$\hat{y}(\mathbf{x}|\mathbf{a}, b) = \arg\min_{y \in \mathbb{Y}} q\big(y,\ z(\mathbf{x}|\mathbf{a}, b)\big)$ решающее правило

Обобщенная линейная модель зависимости

$\begin{cases} z(\mathbf{x}\|\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b: \mathbb{R}^n\to\mathbb{R} \\ q(y,z): \mathbb{Y}\times\mathbb{R}\to\mathbb{R}^+ \end{cases}$	обобщенный линейный признак объекта, представленного вектором признаков $\mathbf{x}\in\mathbb{R}^n$ функция связи, выпуклая по $z\in\mathbb{R}$ для любого $y\in\mathbb{Y}$
$\hat{y}(\mathbf{x}\|\mathbf{a},b)=\arg\min\limits_{y\in\mathbb{Y}} q\big(y, z(\mathbf{x}\|\mathbf{a},b)\big)$	решающее правило

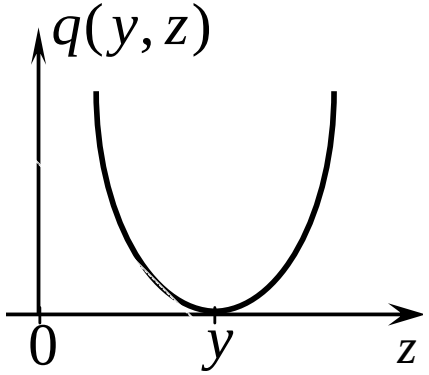
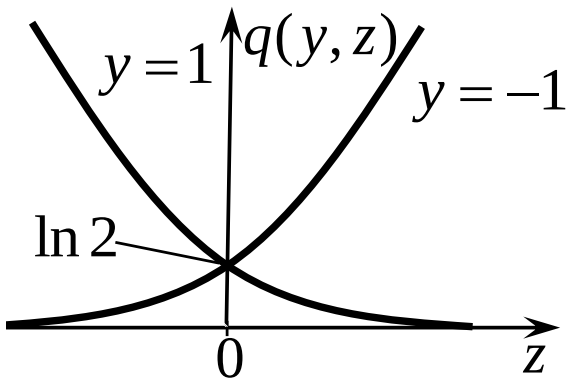
Частные случаи

Регрессия $\mathbb{Y}=\mathbb{R}$	Двух-классовое распознавание образов $\mathbb{Y}=\{-1,1\}$
$q(y,z)=(y-z)^2: \mathbb{R}\times\mathbb{R}\to\mathbb{R}^+$	$\lim\limits_{z\to-\infty} q(y=+1,z)=\infty,\quad \lim\limits_{z\to\infty} q(y=+1,z)=0,$ $\lim\limits_{z\to-\infty} q(y=-1,z)=0,\quad \lim\limits_{z\to\infty} q(y=-1,z)=\infty.$
	

Обобщенная линейная модель зависимости

$\begin{cases} z(\mathbf{x}\|\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b: \mathbb{R}^n\to\mathbb{R} \\ q(y,z): \mathbb{Y}\times\mathbb{R}\to\mathbb{R}^+ \end{cases}$	обобщенный линейный признак объекта, представленного вектором признаков $\mathbf{x}\in\mathbb{R}^n$ функция связи, выпуклая по $z\in\mathbb{R}$ для любого $y\in\mathbb{Y}$
$\hat{y}(\mathbf{x}\|\mathbf{a},b)=\arg\min\limits_{y\in\mathbb{Y}} q\big(y, z(\mathbf{x}\|\mathbf{a},b)\big)$	решающее правило

Частные случаи

Регрессия $\mathbb{Y}=\mathbb{R}$	Двух-классовое распознавание образов $\mathbb{Y}=\{-1,1\}$
$q(y,z)=(y-z)^2: \mathbb{R}\times\mathbb{R}\to\mathbb{R}^+$	$\lim\limits_{z\to-\infty} q(y=+1,z)=\infty,\quad \lim\limits_{z\to\infty} q(y=+1,z)=0,$ $\lim\limits_{z\to-\infty} q(y=-1,z)=0,\quad \lim\limits_{z\to\infty} q(y=-1,z)=\infty.$
	

Логистическая регрессия

$$q(y,z)=\ln\left[1+\exp(-yz)\right]$$

Общепринятый принцип обучения по прецедентам: Минимум регуляризованного эмпирического риска

Множество прецедентов (обучающая совокупность): $\{(\mathbf{x}_j, y_j),\ j=1,...,N\}$

Надо выбрать два параметра $(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R})$ линейной модели

Критерий: Минимум потерь в пределах обучающей совокупности

$EmpR(\mathbf{a},b)=$		$\sum_{j=1}^{N} q(y_j, \mathbf{a}^T\mathbf{x}_j + b)\to\min$	эмпирический риск на обуч. совокуп., вместо среднего риска по всем объектам реального мира

Однако, если $n > N$, то существует континуум точек минимума $(\mathbf{a},b)\in\mathbb{R}^{n+1}$.

Минимум регуляризованного эмпирического риска

$$J(\mathbf{a},b)=\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N} q(y_j, \mathbf{a}^T\mathbf{x}_j+b)\to\min(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R})$$

простейшая ридж регуляризация $\gamma > 0$,

$$J(\mathbf{a},b|\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N} q(y_j, \mathbf{a}^T\mathbf{x}_j+b)\to\min$$

более изощренная селективная регуляризация

Параметр селективности $0\le\mu<\infty$. При увеличении μ, штраф $\mu|a_i|$ «тянет» к нулю коэффициенты при «лишних» признаках, мало способствующих уменьшению эмпирического риска.

Результат – малое подмножество активных признаков: $\hat{\mathbb{I}}(\gamma,\mu)=\{i: a_{\gamma,\mu,i}\neq 0\}\subseteq\{1,...,n\}$

Задача данного доклада: Предложить подходящий способ выбора структурных параметров обобщенной линейной модели зависимости с отбором признаков объектов

Критерий минимума регуляризованного эмпирического риска с отбором признаков содержит два структурных параметра

$$J(\mathbf{a},b|\gamma,\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}q(y_j,\mathbf{a}^T\mathbf{x}_j+b)\to\min(\mathbf{a},b)$$

$\gamma>0$ – ридж коэффициент

$\mu\ge 0$ – коэффициент селективности

Результат работы алгоритма оптимизации (доклады В.В. Моттля и А. Морозова):

$(\hat{\mathbf{a}},\hat{b})_{\gamma,\mu}=\arg\min J(\mathbf{a},b|\gamma,\mu)$ найденные параметры модели

$\hat{\mathbb{I}}_{\gamma,\mu}=\{i:\hat{a}_{\gamma,\mu,i}\ne 0\}\subseteq\{1,\ldots,n\}$ и подмножество активных признаков

Традиционный метод скользящего контроля (Leave-One-Out cross-validation)

Решить задачу обучения N раз без одного обучающего объекта

$$(\hat{\mathbf{a}},\hat{b})^{(j)}_{\gamma,\mu}=\arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{t=1}^{N}q(y_t,\mathbf{a}^T\mathbf{x}_t+b)-q(y_j,\mathbf{a}^T\mathbf{x}_j+b)\right\}$$

Выбрать структурные параметры по критерию $\sum_{t=1}^{N}q(y_t,\hat{\mathbf{a}}^{(j)T}_{\gamma,\mu}\mathbf{x}_t+\hat{b}^{(j)}_{\gamma,\mu})\to\min_{\gamma,\mu}$

Недостаток – вычисление критерия требует N-кратного обучения

Традиционный метод скользящего контроля (Leave-One-Out Cross-Validation)

Решить задачу обучения N раз без одного обучающего объекта

$$(\hat{\mathbf{a}},\hat{b})_{\gamma,\mu}^{(j)} = \arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|,\ |a_i|\le\mu\\ \mu^2+a_i^2,\ |a_i|>\mu\end{pmatrix}+\sum_{t=1}^{N}q(y_t,\mathbf{a}^T\mathbf{x}_t+b)-q(y_j,\mathbf{a}^T\mathbf{x}_j+b)\right\}$$

Выбрать структурные параметры по критерию $\sum_{t=1}^{N}q(y_t,\hat{\mathbf{a}}_{\gamma,\mu}^{(j)T}\mathbf{x}_t+\hat{b}_{\gamma,\mu}^{(j)})\to\min_{\gamma,\mu}$

Недостаток – вычисление критерия требует N-кратного обучения

Наша идея – принцип дифференциального скользящего контроля (Differential LOO)

1) Удалять из обучающей совокупности не весь очередной объект, а лишь малую его часть, не приводящую к изменению подмножества активных признаков

$$(\hat{\mathbf{a}},\hat{b})_{\gamma,\mu}(p_j) = \arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|,\ |a_i|\le\mu\\ \mu^2+a_i^2,\ |a_i|>\mu\end{pmatrix}+\sum_{t=1}^{N}q(y_t,\mathbf{a}^T\mathbf{x}_t+b)-p_jq(y_j,\mathbf{a}^T\mathbf{x}_j+b)\right\},\ p_j\to 0$$

$$\hat{\mathbb{I}}_{\gamma,\mu}(p_j)=\left\{i: a_{\gamma,\mu,i}^{(p_j)}\neq 0\right\}\subseteq\{1,\ldots,n\}=const$$

2) В качестве критерия принять сумму производных потерь на всех объектах

$DLOO(\gamma,\mu)=\sum_{j=1}^{N}\frac{\partial}{\partial p_j}q\left(y_j,\mathbf{x}_j^T\hat{\mathbf{a}}_{\gamma,\mu}(p_j)+\hat{b}_{\gamma,\mu}(p_j)\right)$. Такой критерий показывает сумму скоростей роста потерь при «тенденции к удалению» каждого объекта из обучения.

Наша идея – принцип дифференциального скользящего контроля (Differential LOO, DLOO)

1) Удалять из обучающей совокупности не весь очередной объект, а лишь малую его часть, не приводящую к изменению подмножества активных признаков

$$(\hat{\mathbf{a}},\hat{b})_{\gamma,\mu}(p_j)=\arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|,\ |a_i|\le\mu\\ \mu^2+a_i^2,\ |a_i|>\mu\end{pmatrix}+\sum_{t=1}^{N}q(y_t,\mathbf{a}^T\mathbf{x}_t+b)-p_jq(y_j,\mathbf{a}^T\mathbf{x}_j+b)\right\},\ p_j\to 0$$

$$\hat{\mathbb{I}}_{\gamma,\mu}(p_j)=\left\{i:\hat{a}_{\gamma,\mu,i}(p_j)\neq 0\right\}\subseteq\{1,\ldots,n\}=const$$

2) В качестве критерия принять сумму производных потерь на всех объектах

$DLOO(\gamma,\mu)=\sum_{j=1}^{N}\frac{\partial}{\partial p_j}q\left(y_j,\mathbf{x}_j^T\hat{\mathbf{a}}_{\gamma,\mu}(p_j)+\hat{b}_{\gamma,\mu}(p_j)\right)$. Такой критерий показывает сумму скоростей роста потерь при «тенденции к удалению» каждого объекта из обучения.

В данном докладе мы применим эту идею к частному случаю селективной модели регрессионной зависимости

Квадратичная функция связи функция $q\left(y_t,\mathbf{a}^T\mathbf{x}_t+b\right)=\left(y_t-(\mathbf{a}^T\mathbf{x}_t+b)\right)^2$

Для упрощения выкладок будем полагать $b=0$. Исходная задача:

$$J(\mathbf{a}|\gamma,\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|,\ |a_i|\le\mu\\ \mu^2+a_i^2,\ |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}(y_j-\mathbf{a}^T\mathbf{x}_j)^2\to\min(\mathbf{a}),\ \ \mathbf{x}_j,\mathbf{a}\in\mathbb{R}^n$$

В данном докладе мы применим эту идею к частному случаю селективной модели регрессионной зависимости

Квадратичная функция связи $q\left(y_t, \mathbf{x}_t^T\mathbf{a}+b\right)=\left(y_t-(\mathbf{x}_t^T\mathbf{a}+b)\right)^2$

Для упрощения выкладок будем полагать $b=0$. Исходная задача:

$$J(\mathbf{a}|\gamma,\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2\to\min(\mathbf{a}),\ \ \mathbf{x}_t,\mathbf{a}\in\mathbb{R}^n$$

Сначала рассмотрим задачу без селективности

$$\hat{\mathbf{a}}_\gamma=\arg\min\left\{\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N}(y_j-\mathbf{a}^T\mathbf{x}_j)^2\right\}.$$

$$\hat{\mathbf{a}}_\gamma^{(j)}=\arg\min\left\{\gamma\,\mathbf{a}^T\,\mathbf{a}+\sum_{t=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2-(y_j-\mathbf{x}_j^T\mathbf{a})^2\right\}$$

удалим из обучения один объект j

Классический критерий *LOO*

$$Loo(\gamma)=\sum_{j=1}^{N}\left(\qquad y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_\gamma^{(j)}\qquad\right)^2$$

В данном докладе мы применим эту идею к частному случаю селективной модели регрессионной зависимости

Квадратичная функция связи $q\left(y_t, \mathbf{x}_t^T\mathbf{a}+b\right)=\left(y_t-(\mathbf{x}_t^T\mathbf{a}+b)\right)^2$

Для упрощения выкладок будем полагать $b=0$. Исходная задача:

$$J(\mathbf{a}|\gamma,\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2\to\min(\mathbf{a}),\ \ \mathbf{x}_t,\mathbf{a}\in\mathbb{R}^n$$

Сначала рассмотрим задачу без селективности

$$\hat{\mathbf{a}}_\gamma=\arg\min\left\{\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N}(y_j-\mathbf{a}^T\mathbf{x}_j)^2\right\}.$$

$$\hat{\mathbf{a}}_\gamma^{(j)}=\arg\min\left\{\gamma\,\mathbf{a}^T\,\mathbf{a}+\sum_{t=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2-(y_j-\mathbf{x}_j^T\mathbf{a})^2\right\}$$

удалим из обучения один объект j

Классический критерий *LOO* – перебор не требуется

$$Loo(\gamma)=\sum_{j=1}^{N}\left(\frac{y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_\gamma}{1-\mathbf{x}_j^T\left((\mathbf{X}\mathbf{X}^T)+\gamma\mathbf{I}_n\right)^{-1}\mathbf{x}_j}\right)^2$$

M.A. Goldberg, H.A. Cho. Introduction to Regression Analysis. WIT Press, 2004. The Sherman-Morrison-Woodbury formula. Page 141.

В данном докладе мы применим эту идею к частному случаю селективной модели регрессионной зависимости

Квадратичная функция связи $q\left(y_t, \mathbf{x}_t^T\mathbf{a}+b\right)=\left(y_t-(\mathbf{x}_t^T\mathbf{a}+b)\right)^2$

Для упрощения выкладок будем полагать $b=0$. Исходная задача:

$$J(\mathbf{a}|\gamma,\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2\to\min(\mathbf{a}),\ \ \mathbf{x}_t,\mathbf{a}\in\mathbb{R}^n$$

Сначала рассмотрим задачу без селективности

$$\hat{\mathbf{a}}_\gamma=\arg\min\left\{\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N}(y_j-\mathbf{a}^T\mathbf{x}_j)^2\right\}.$$

$$\hat{\mathbf{a}}_\gamma^{(j)}=\arg\min\left\{\gamma\,\mathbf{a}^T\,\mathbf{a}+\sum_{t=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2-(y_j-\mathbf{x}_j^T\mathbf{a})^2\right\}$$

удалим из обучения один объект j

Классический критерий *LOO* – перебор не требуется

$$Loo(\gamma)=\sum_{j=1}^{N}\left(\frac{y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_\gamma}{1-\mathbf{x}_j^T\left((\mathbf{X}\mathbf{X}^T)+\gamma\mathbf{I}_n\right)^{-1}\mathbf{x}_j}\right)^2$$

Но мы не можем так делать из-за присутствия селективной регуляризации, изменяющей размерность векторов признаков.

Сначала рассмотрим задачу без селективности

$$\hat{\mathbf{a}}_\gamma^{(j)} = \arg\min\left\{\gamma\,\mathbf{a}^T\,\mathbf{a} + \sum_{t=1}^{N}(y_t - \mathbf{x}_t^T\mathbf{a})^2 - (y_j - \mathbf{x}_j^T\mathbf{a})^2\right\}$$

удалим из обучения один объект j

Классический критерий *LOO* – перебор не требуется

$$Loo(\gamma) = \sum_{j=1}^{N}\left(\frac{y_j - \mathbf{x}_j^T\hat{\mathbf{a}}_\gamma}{1 - \mathbf{x}_j^T\left((\mathbf{X}\mathbf{X}^T) + \gamma\mathbf{I}_n\right)^{-1}\mathbf{x}_j}\right)^2$$

Но мы не можем так делать из-за присутствия селективной регуляризации, изменяющей размерность векторов признаков.

$$\hat{\mathbf{a}}_\gamma^{(j)} = \arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2 + a_i^2, & |a_i|>\mu\end{pmatrix} + \sum_{j=1}^{N}(y_t - \mathbf{x}_t^T\mathbf{a})^2 - (y_j - \mathbf{x}_j^T\mathbf{a})^2\right\}$$

Обычный LOO: полное удаление

Вместо полного удаления очередного объекта удалим его малую часть, тогда оценка направляющего вектора изменится мало, и его размерность не изменится.

$$\hat{\mathbf{a}}_\gamma^{(p_j)} = \arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2 + a_i^2, & |a_i|>\mu\end{pmatrix} + \sum_{j=1}^{N}(y_t - \mathbf{x}_t^T\mathbf{a})^2 - p_j(y_j - \mathbf{x}_j^T\mathbf{a})^2\right\}$$

Дифференц. LOO: $p_j \to 0$

Остается вычислить скорости роста ошибок $\partial/\partial p_j\left(y_j - \mathbf{x}_j^T\hat{\mathbf{a}}_\gamma^{(p_j)}\right)$.

Дифференциальный LOO для квадратичной модели

$$\hat{\mathbf{a}}_\gamma(p_j)=\arg\min\left\{\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2-p_j(y_j-\mathbf{x}_j^T\mathbf{a})^2\right\}$$ Дифференц. LOO: $p_j\to 0$

$\hat{\mathbf{a}}_\gamma=\hat{\mathbf{a}}_\gamma(0)$ – оценка направляющего вектора по всей обучающей совокупности

$\hat{\mathbf{a}}_\gamma^{(j)}=\hat{\mathbf{a}}_\gamma(1)$ – полный LOO, $p_j=1$, $\hat{\mathbf{a}}_\gamma(p_j)$ – зависимость от удаляемой доли объекта

$\boxed{\hat{\mathbf{a}}_\gamma(p_j)=(1-p_j)\hat{\mathbf{a}}_\gamma+p_j\hat{\mathbf{a}}_\gamma^{(j)}}$ производная $\dfrac{\partial}{\partial p_j}\hat{\mathbf{a}}_\gamma(p_j)=\hat{\mathbf{a}}_\gamma^{(j)}-\hat{\mathbf{a}}_\gamma$

Теорема. В этом случае $\boxed{\dfrac{\partial}{\partial p_j}\left(y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_\gamma(p_j)\right)=\dfrac{y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_\gamma}{1-\mathbf{x}_j^T\left((\mathbf{X}\mathbf{X}^T)+\gamma\mathbf{I}_n\right)^{-1}\mathbf{x}_j}}$

Итоговый дифференциальный критерий LOO для квадратичной модели формально совпадает с обычным критерием LOO

$$Dif\,Loo(\gamma)=\sum_{j=1}^{N}\left(\frac{y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_\gamma}{1-\mathbf{x}_j^T\left((\mathbf{X}\mathbf{X}^T)+\gamma\mathbf{I}_n\right)^{-1}\mathbf{x}_j}\right)^2$$

Вопрос: Как применить этот критерий к селективной модели?

$$J(\mathbf{a}|\gamma,\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|,&|a_i|\le\mu\\ \mu^2+a_i^2,&|a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}(y_t-\mathbf{x}_t^T\mathbf{a})^2\to\min(\mathbf{a})$$

Дифференциальный LOO для селективной квадратичной модели

Пусть задача решена при конкретных значениях структурных параметров (γ,μ)

$$\hat{\mathbf{a}}_{\gamma,\mu}=\arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|,\ |a_i|\le\mu\\ \mu^2+a_i^2,\ |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}\left(y_t-\sum_{i=1}^{n}a_i x_{j,i}\right)^2\right\}$$

$\hat{\mathbf{a}}_{\gamma,\mu}=\left(\hat{a}_{\gamma,\mu,1}\cdots\hat{a}_{\gamma,\mu,n}\right)$ доклады В.В. Моттля и А.О. Морозова

$\hat{\mathbb{I}}_{\gamma,\mu}=\left\{i:\hat{a}_{\gamma,\mu,i}\neq 0\right\}\subseteq\{1,\ldots,n\}$ – подмножество $\hat{n}_{\gamma,\mu}=\left|\hat{\mathbb{I}}_{\gamma,\mu}\right|\le n$ активных признаков

$\mathbf{x}_{\gamma,\mu,j}=\left(x_{j,i},\ i\in\hat{\mathbb{I}}_{\gamma,\mu}\right)\in\mathbb{R}^{\hat{n}_{\gamma,\mu}}$ – активные признаки $\quad\mathbf{X}_{\gamma,\mu}=\left(\mathbf{x}_{\gamma,\mu,1}\cdots\mathbf{x}_{\gamma,\mu,n}\right)$ – матрица $(\hat{n}_{\gamma,\mu}\times N)$

Свойство решения:

$$\hat{\mathbf{a}}_{\gamma,\mu}=\arg\min\left\{\gamma\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}a_i^2+\sum_{j=1}^{N}\left(y_t-\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}a_i x_{j,i}\right)^2\right\}$$

Если запомнить подмножество активных признаков $\hat{\mathbb{I}}_{\gamma,\mu}=\left\{i:\hat{a}_{\gamma,\mu,i}\neq 0\right\}$, то вектор $\hat{\mathbf{a}}_{\gamma,\mu}$ легко восстановить

Это эквивалентная модель зависимости, в которой роль структурных параметров вместо пары (γ,μ) играет пара $\left(\mathbb{I}_{\gamma,\mu},\gamma\right)$. Идея дифференциального критерия LOO для исходной селективной модели выражается условием

$$Dif\,Loo(\gamma,\mu)=\sum_{j=1}^{N}\left(\frac{y_j-\mathbf{x}^T_{\hat{\mathbb{I}}_{\gamma,\mu},j}\hat{\mathbf{a}}_{\hat{\mathbb{I}}_{\gamma,\mu},\gamma}}{1-\mathbf{x}^T_{\hat{\mathbb{I}}_{\gamma,\mu},j}\left((\mathbf{X}_{\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{X}^T_{\hat{\mathbb{I}}_{\gamma,\mu}})+\gamma\mathbf{I}_{\hat{n}_{\gamma,\mu}}\right)^{-1}\mathbf{x}_{\hat{\mathbb{I}}_{\gamma,\mu},j}}\right)^2\to\min(\gamma,\mu)$$

Эксперимент

Регрессоры $\mathbf{x}_t = (x_{t,i}, i = 1,\ldots,n), t = 1,\ldots,N = 251$:

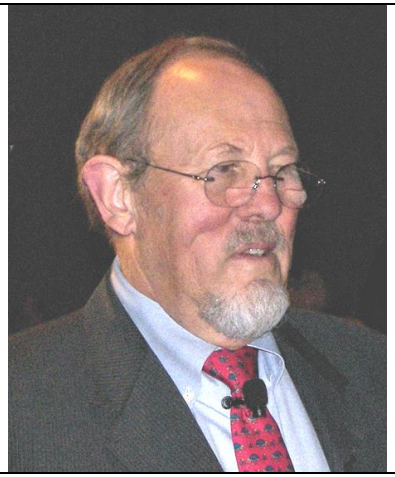
Множество временных рядов месячных доходностей $n = 650$ биржевых ценных бумаг на Нью-Йоркской фондовой бирже чуть больше, чем за два года.

Наблюдаемый сигнал $y_t, t = 1,\ldots,N = 251$:

Временной ряд доходностей инвестиционного портфеля, построенного как вложение капитала в равных долях в $n^* = 13$ неизвестных ценных бумаг в множестве $\{1,\ldots,n\}$.

Требуется найти подмножество активных ценных бумаг.

Регрессионная модель: $y_t \cong \sum_{i=1}^{n} a_i x_{t,i}$ – модель Шарпа, a_i – доли вложения капитала.



Уильям Шарп, почетныйProfessor of Finance
Graduate School of Business, Stanford University

Лауреат Нобелевской премии по экономике 1990 г.
за ряд достижений, в частности, за
Returns Based Style Analysis –RBSA
оценивание стиля и эффективности инвестиционных фондов

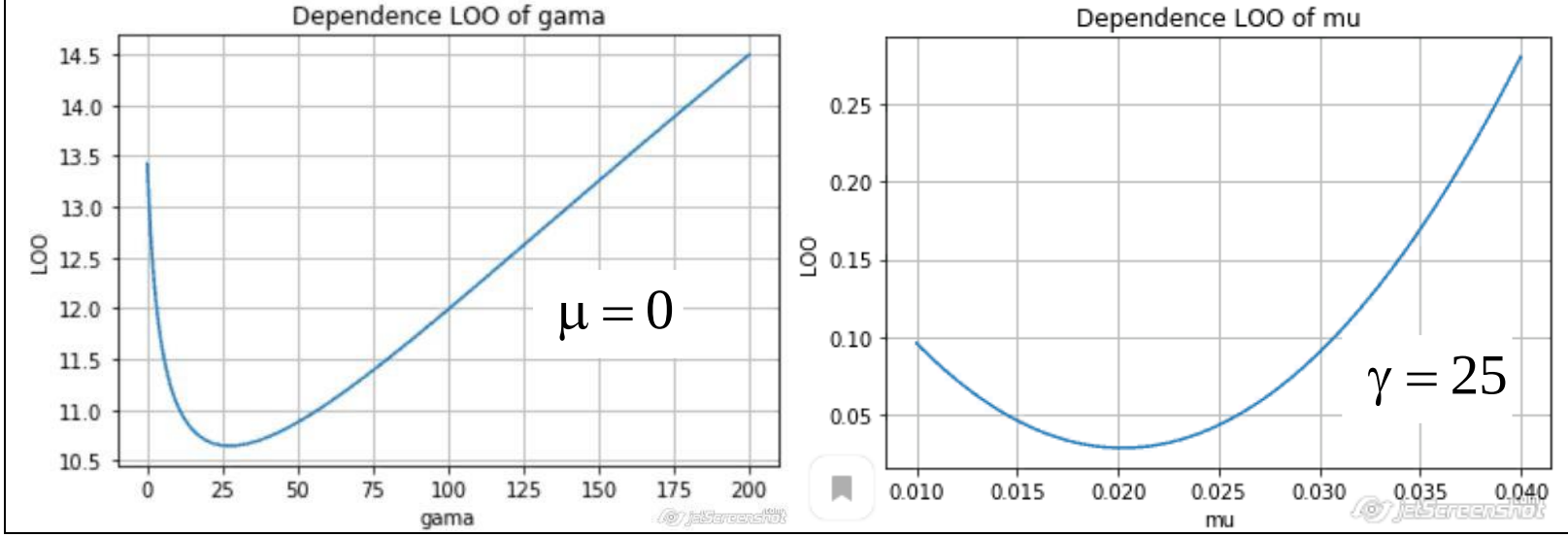
Мы рассматриваем дополнительную задачу Factor Search – оценивание фактического состава инвестиционного портфеля в большом множестве биржевых активов. У Шарпа задачи Factor Search не было.

Результат идентификации селективной модели

$$\hat{\mathbf{a}}_{\gamma,\mu} = \arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix} 2\mu|a_i|, & |a_i|\le\mu \\ \mu^2+a_i^2, & |a_i|>\mu \end{pmatrix}+\sum_{j=1}^{N}\left(y_t-\sum_{i=1}^{n}a_i x_{j,i}\right)^2\right\}, \; n=650, \;\; N=251,$$

по критерию дифференциального скользящего контроля

$$Dif\,Loo(\gamma,\mu)=\sum_{j=1}^{N}\left(\frac{y_j-\mathbf{x}^T_{\hat{\mathbb{I}}_{\gamma,\mu},j}\hat{\mathbf{a}}_{\hat{\mathbb{I}}_{\gamma,\mu},\gamma}}{1-\mathbf{x}^T_{\hat{\mathbb{I}}_{\gamma,\mu},j}\left((\mathbf{X}_{\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{X}^T_{\hat{\mathbb{I}}_{\gamma,\mu}})+\gamma\mathbf{I}_{\hat{n}_{\gamma,\mu}}\right)^{-1}\mathbf{x}_{\hat{\mathbb{I}}_{\gamma,\mu},j}}\right)^2\to\min(\gamma,\mu)$$



Оценка:
$\hat{\gamma}=25,\ \hat{\mu}=0,02$

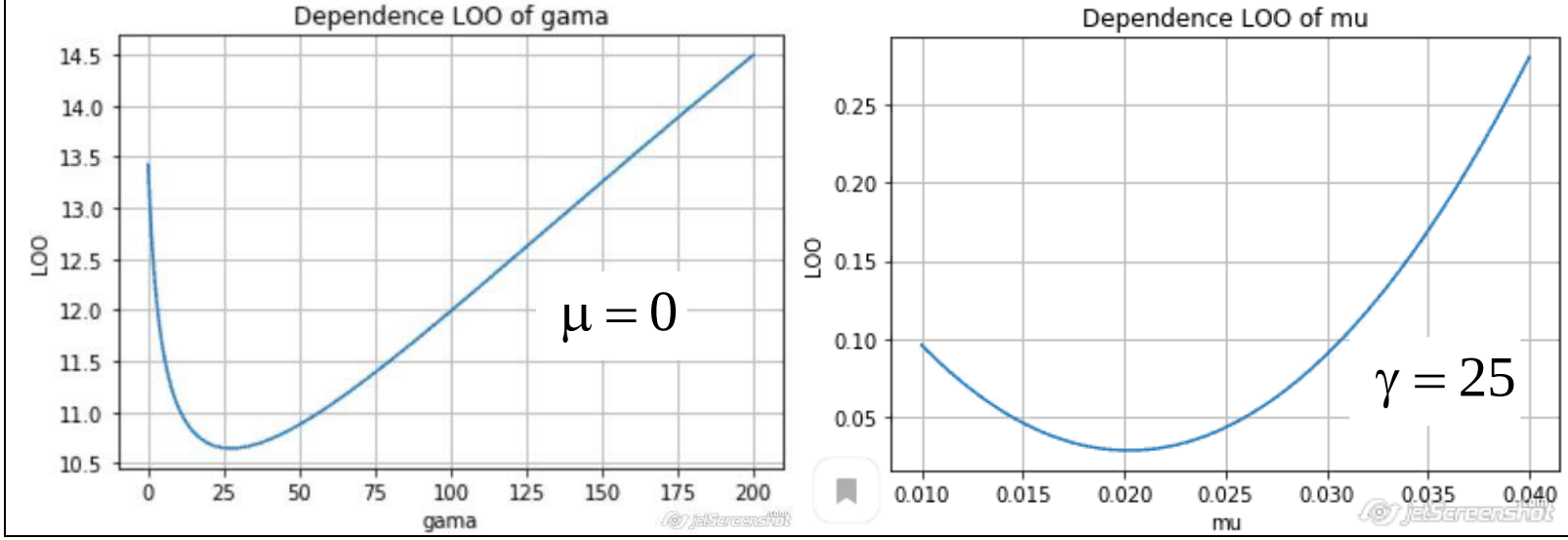
Найденное подмножество $\hat{\mathbb{I}}_{\hat{\gamma},\hat{\mu}}=\left\{i:\hat{a}_{\hat{\gamma},\hat{\mu},i}\neq 0\right\}$ содержит $\hat{n}_{\hat{\gamma},\hat{\mu}}=16$ биржевых активов, содержащее n^*=13 истинных.

Результат идентификации селективной модели

$$\hat{\mathbf{a}}_{\gamma,\mu} = \arg\min\left\{\gamma\sum_{i=1}^{n}\begin{pmatrix} 2\mu|a_i|, & |a_i|\le\mu \\ \mu^2+a_i^2, & |a_i|>\mu \end{pmatrix}+\sum_{j=1}^{N}\left(y_t-\sum_{i=1}^{n}a_i x_{j,i}\right)^2\right\}, \; n=650, \;\; N=251,$$

по критерию дифференциального скользящего контроля

$$Dif\,Loo(\gamma,\mu)=\sum_{j=1}^{N}\left(\frac{y_j-\mathbf{x}^T_{\hat{\mathbb{I}}_{\gamma,\mu},j}\hat{\mathbf{a}}_{\hat{\mathbb{I}}_{\gamma,\mu},\gamma}}{1-\mathbf{x}^T_{\hat{\mathbb{I}}_{\gamma,\mu},j}\left((\mathbf{X}_{\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{X}^T_{\hat{\mathbb{I}}_{\gamma,\mu}})+\gamma\mathbf{I}_{\hat{n}_{\gamma,\mu}}\right)^{-1}\mathbf{x}_{\hat{\mathbb{I}}_{\gamma,\mu},j}}\right)^2\to\min(\gamma,\mu)$$



Оценка:
$\hat{\gamma}=25,\ \hat{\mu}=0,02$

Найденное подмножество $\hat{\mathbb{I}}_{\hat{\gamma},\hat{\mu}}=\left\{i:\hat{a}_{\hat{\gamma},\hat{\mu},i}\neq 0\right\}$ содержит $\hat{n}_{\hat{\gamma},\hat{\mu}}=16$ биржевых активов, содержащее n^*=13 истинных.

Можно сказать, что мы нашли иголку в стоге сена!

Благодарности

Работа поддержана грантами РФФИ:
19-37-90159-Аспиранты (А.О. Морозов),
17-07-00436-а (В.В. Моттль).

Спасибо за внимание!

Вопросы?