Проверка обоснованности обучаемых моделей зависимостей. Обобщенный линейный подход

Моттль Вадим Вячеславович
Вычислительный центр РАН

Левдик Павел Владимирович
Московский физико-технический институт

Красоткина Ольга Вячеславовна
Московский государственный университет

Типовая задача восстановления закономерности на множестве объектов реального мира

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

Некоторое множество значений скрытой характеристики объектов $y \in \mathbb{Y}$.

Типовая задача восстановления закономерности на множестве объектов реального мира

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

Некоторое множество значений скрытой характеристики объектов $y \in \mathbb{Y}$.

Природа выбирает пару $(\omega, y) \in \Omega \times \mathbb{Y}$, где наблюдателю недоступно значение $y \in \mathbb{Y}$

Желание наблюдателя:

Иметь инструмент оценивания скрытой характеристики для реальных объектов

$\hat{y}(\omega) : \Omega \to \mathbb{Y}$; $\hat{y}(\omega) \neq y(\omega)$ – ошибка.

Типовая задача восстановления закономерности на множестве объектов реального мира

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

Некоторое множество значений скрытой характеристики объектов $y \in \mathbb{Y}$.

Природа выбирает пару $(\omega, y) \in \Omega \times \mathbb{Y}$, где наблюдателю недоступно значение $y \in \mathbb{Y}$

Желание наблюдателя:

Иметь инструмент оценивания скрытой характеристики для реальных объектов

$\hat{y}(\omega) : \Omega \to \mathbb{Y}$; $\hat{y}(\omega) \neq y(\omega)$ – ошибка.

Обучение по прецедентам:

Подмножество наблюдаемых объектов $\Omega^* \subset \Omega$, для которых измерено значение целевой характеристики $\left(\omega \in \Omega^*, y\right)$:

$\left\{(\omega_j, y_j),\ j = 1, ..., N\right\}$ – обучающая совокупность.

Типовая задача восстановления закономерности на множестве объектов реального мира

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

Некоторое множество значений скрытой характеристики объектов $y \in \mathbb{Y}$.

Природа выбирает пару $(\omega, y) \in \Omega \times \mathbb{Y}$, где наблюдателю недоступно значение $y \in \mathbb{Y}$

Желание наблюдателя:

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

Обучение по прецедентам:

Подмножество наблюдаемых объектов $\Omega^* \subset \Omega$, для которых измерено значение целевой характеристики $\left(\omega \in \Omega^*, y\right)$:

$\left\{(\omega_j, y_j),\ j = 1, \ldots, N\right\}$ – обучающая совокупность.

Задача: Выбрать функцию $\hat{y}(\omega)$, определенную на всем множестве Ω, так чтобы можно было в дальнейшем оценивать значение рассматриваемой характеристики для новых объектов $\omega \in \Omega \setminus \Omega^*$.

Типовая задача восстановления закономерности на множестве объектов реального мира

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

Некоторое множество значений скрытой характеристики объектов $y \in \mathbb{Y}$.

Природа выбирает пару $(\omega, y) \in \Omega \times \mathbb{Y}$, где наблюдателю недоступно значение $y \in \mathbb{Y}$

Желание наблюдателя:

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

Типовые случаи:

Задача обучения распознаванию образов
$\mathbb{Y} = \{y_1, ..., y_m\}$ – конечное неупорядоченное множество; в частности $\mathbb{Y} = \{-1, 1\}$.

Задача восстановления числовой регрессии, т.е. оценивания числовой функции
$\mathbb{Y} = \mathbb{R}$ – множество действительных чисел.

Задача восстановления ранговой регрессии
$\mathbb{Y} = \{y_1 < ... < y_m\}$ – конечное множество с отношением линейного порядка.

Представление объектов в виде, допускающем ввод данных в компьютер

Непосредственно построить оценочную функцию $\hat{y}(\omega)$ невозможно, поскольку объект реального мира $\omega \in \Omega$ не может быть явно представлен в компьютере.

1. Индивидуальное представление объектов.

$\boldsymbol{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega) \right)^T : \Omega \to \mathbb{R}^n$	Простейшая ситуация: вектор из n числовых признаков

Обучающая совокупность: $\left\{ \left(\boldsymbol{x}_j = \boldsymbol{x}(\omega_j),\ y_j = y(\omega_j) \right),\ j = 1, \ldots, N \right\}$.

2. Сравнительное числовое представление объектов.

Объекты могут быть восприняты компьютером только через их попарное сравнение.
$S(\omega', \omega''): \Omega \times \Omega \to \mathbb{R}$ – числовая функция, выражающая некоторое сравнительное свойство пар объектов.

Базисная совокупность $\left\{ \omega_i^0, i = 1, \ldots, n \right\} \subset \Omega$	Нет значений целевой характеристики, есть только матрица парных сравнений $\left[S(\omega_i, \omega_k),\ i, k = 1, \ldots, n \right]$
$\boldsymbol{x}(\omega) = \left(x_i(\omega) = S(\omega_i, \omega),\ i = 1, \ldots, n \right) \in \mathbb{R}^n$	Вектор вторичных признаков объекта $\omega \in \Omega$ относительно базисной совокупности
Обучающая совокупность (часть базисной совокупности) $\{ \omega_j,\ j = 1, \ldots, N \} \subset \Omega$	Известны значения целевой характеристики $\left\{ \left(\boldsymbol{x}_j = \boldsymbol{x}(\omega_j),\ y_j = y(\omega_j) \right),\ j = 1, \ldots, N \right\}$

Обобщенная линейная модель зависимости

Вектор вторичных признаков объекта
$\boldsymbol{x}(\omega)=\big(x_i(\omega)=S(\omega_i,\omega),\ i=1,...,n\big)\in\mathbb{R}^n$

Функция парного сравнения
$\boxed{S(\omega',\omega''): \Omega\times\Omega\to\mathbb{R}\ \text{– фантазия наблюдателя}}$

Параметры модели зависимости $(\boldsymbol{a},b)$
$\begin{cases}\boldsymbol{a}\in\mathbb{R}^n & \text{– направляющий вектор}\\ b\in\mathbb{R} & \text{– сдвиг модели}\end{cases}$

Обобщенный линейный признак объекта $z(\boldsymbol{x},\boldsymbol{a},b)=\boldsymbol{a}^T\boldsymbol{x}+b:\ \mathbb{R}^n\xrightarrow{(\boldsymbol{a},b)}\mathbb{R}$

Целевая характеристика объекта $y(\omega)\in\mathbb{Y}$ – задана природой

Функция связи (link function), обычно выпуклая по z
$\boxed{q(y,z): \mathbb{Y}\times\mathbb{R}\to\mathbb{R}^+\ \text{– фантазия наблюдателя}}$

Параметрическая функция потерь $q(y,\boldsymbol{x},\boldsymbol{a},b)=q\big(y,z(\boldsymbol{x},\boldsymbol{a},b)\big): \mathbb{Y}\times\mathbb{R}^n\xrightarrow{(\boldsymbol{a},b)}\mathbb{R}^+$

Решающее правило $\hat{y}(\boldsymbol{x}\,|\,\boldsymbol{a},b)=\operatorname*{argmin}_{y\in\mathbb{Y}} q(y,\boldsymbol{x},\boldsymbol{a},b):\ \mathbb{X}\xrightarrow{(\boldsymbol{a},b)}\mathbb{Y}$

Обучающая совокупность $(\boldsymbol{X},\boldsymbol{Y})=\Big\{\big(\boldsymbol{x}(\omega_j),y(\omega_j)\big)=(\boldsymbol{x}_j,y_j),\ j=1,...,N\Big\}$

Эмпирический риск
$$Q(\boldsymbol{v},b\,|\,X,Y)=\frac{1}{N}\sum_{j=1}^N q(y_j,\boldsymbol{x}_j,\boldsymbol{v},b)$$

Обучение – выбор $\big(\boldsymbol{a}\in\mathbb{R}^n,b\in\mathbb{R}\big)$**:**
Минимизация эмпирического риска

$$\sum_{j=1}^N q\big(y_j,z(\boldsymbol{x}_j,\boldsymbol{a},b)\big)\to\min(\boldsymbol{a},b),$$
$$z(\boldsymbol{x}_j,\boldsymbol{a},b)=\boldsymbol{a}^T\boldsymbol{x}+b$$

Обобщенная линейная модель зависимости

Вектор вторичных признаков объекта $\boldsymbol{x}(\omega)=\big(x_i(\omega)=S(\omega_i,\omega\,|\,\alpha),\ i=1,...,n\big)\in\mathbb{R}^n$

Параметрическое семейство функций сравнения $\boxed{S(\omega',\omega''\,|\,\alpha):\Omega\times\Omega\to\mathbb{R}\ \text{– фантазия наблюдателя}}$

Параметры модели зависимости $(\boldsymbol{a},b)$ $\begin{cases}\boldsymbol{a}\in\mathbb{R}^n\ \text{– направляющий вектор}\\ b\in\mathbb{R}\ \text{– сдвиг модели}\end{cases}$

Обобщенный линейный признак объекта $z(\boldsymbol{x},\boldsymbol{a},b)=\boldsymbol{a}^T\boldsymbol{x}+b:\ \mathbb{R}^n\xrightarrow{(\boldsymbol{a},b)}\mathbb{R}$

Целевая характеристика объекта $y(\omega)\in\mathbb{Y}$ – задана природой

Функция связи (link function), обычно выпуклая по z $\boxed{q(y,z):\mathbb{Y}\times\mathbb{R}\to\mathbb{R}^+\ \text{– фантазия наблюдателя}}$

Параметрическая функция потерь $q(y,\boldsymbol{x},\boldsymbol{a},b)=q\big(y,z(\boldsymbol{x},\boldsymbol{a},b)\big):\mathbb{Y}\times\mathbb{R}^n\xrightarrow{(\boldsymbol{a},b)}\mathbb{R}^+$

Решающее правило $\hat{y}(\boldsymbol{x}\,|\,\boldsymbol{a},b)=\underset{y\in\mathbb{Y}}{\operatorname{argmin}}\,q(y,\boldsymbol{x},\boldsymbol{a},b):\ \mathbb{X}\xrightarrow{(\boldsymbol{a},b)}\mathbb{Y}$

Обучающая совокупность $(\boldsymbol{X},\boldsymbol{Y})=\Big\{\big(\boldsymbol{x}(\omega_j),y(\omega_j)\big)=(\boldsymbol{x}_j,y_j),\ j=1,...,N\Big\}$

Семейство выпуклых регуляризующих функций $\boxed{V(\boldsymbol{a}\,|\,\mu):\ \mathbb{R}^n\xrightarrow{\mu}\mathbb{R}^+\ \text{– еще одна фантазия наблюдателя}}$

Обучение – выбор $\big(\boldsymbol{a}\in\mathbb{R}^n,b\in\mathbb{R}\big)$: Минимизация регуляризованного эмпирического риска

$$V(\boldsymbol{a}\,|\,\mu)+c\sum_{j=1}^{N}q\Big(y_j,z(\boldsymbol{x}_j,\boldsymbol{a},b)\Big)\to\min(v,b),$$
$$z(\boldsymbol{x}_j,\boldsymbol{a},b)=K(\boldsymbol{x}_j,v)+b$$

Критерий выпуклый, если регуляризующая функция $V(\boldsymbol{a}\,|\,\mu)$ и функция связи $q(y,z)$ выпуклы по $\boldsymbol{a}\in\mathbb{R}^n$ и $z\in\mathbb{R}$.

Вероятностная интерпретация минимизации регуляризованного эмпирического риска

Обучение: Минимизация регуляризованного эмпирического риска на обучающей совокупности

$$(X,Y)=\Big\{\big(x(\omega_j),y(\omega_j)\big)=(x_j,y_j),\ j=1,...,N\Big\}$$

$$V(\boldsymbol{a}\,|\,\mu)+c\sum_{j=1}^{N}q\big(y_j,z(\boldsymbol{x}_j,\boldsymbol{a},b)\big)\to\min(\boldsymbol{a},b),$$

$$z(\boldsymbol{x}_j,\boldsymbol{a},b)=K(\boldsymbol{x}_j,\boldsymbol{a})+b$$

Регуляризующая функция $V(\boldsymbol{a}|\mu)$ и функция связи $q(y,z)$ выпуклы по $\boldsymbol{a}\in\mathbb{R}^n$ и $z\in\mathbb{R}$.

Вероятностные эвристики наблюдателя

1) Чем меньше параметрическая функция потерь $q\big(y,z(\boldsymbol{x},\boldsymbol{a},b)\big)$, тем более вероятно появление объекта с характеристиками $(\boldsymbol{x},y)$.

Предполагаемое параметрическое семейство совместных распределений $\varphi(\boldsymbol{x},y\,|\,\boldsymbol{a},b,c)\propto\exp\Big\{-cq\big(y,z(\boldsymbol{x},\boldsymbol{a},b)\big)\Big\}$.

2) Чем меньше регуляризующая функция $V(\boldsymbol{a}|\mu)$, тем более априори вероятно значение параметра $\boldsymbol{a}\in\mathbb{R}^n$; предположений о b нет.

Предполагаемое параметрическое семейство априорных распределений $\psi(\boldsymbol{a},b|\mu)\propto\psi(\boldsymbol{a}|\mu)\propto\exp\big\{-c'V(\boldsymbol{a}|\mu)\big\}$.

Байесовский принцип обучения: максимизации апостериорной плотности на обучающей совокупности

$$(\boldsymbol{X},\boldsymbol{Y})=\Big\{\big(\boldsymbol{x}(\omega_j),y(\omega_j)\big)=(\boldsymbol{x}_j,y_j),\ j=1,...,N\Big\}$$

$$P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)\propto$$

$$\psi(\boldsymbol{a}|\mu)\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\to\max(\boldsymbol{a},b),$$

$$\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)=\prod_{j=1}^{N}\varphi(\boldsymbol{x}_j,y_j\,|\,\boldsymbol{a},b,c).$$

Вероятностная интерпретация минимизации регуляризованного эмпирического риска

Обучение: Минимизация регуляризованного эмпирического риска на обучающей совокупности

$$(X,Y)=\Big\{\big(x(\omega_j),y(\omega_j)\big)=(x_j,y_j),\ j=1,...,N\Big\}$$

$$V(\boldsymbol{a}\,|\,\mu)+c\sum_{j=1}^{N}q\big(y_j,z(\boldsymbol{x}_j,\boldsymbol{a},b)\big)\to\min(\boldsymbol{a},b),$$

$$z(\boldsymbol{x}_j,\boldsymbol{a},b)=K(\boldsymbol{x}_j,\boldsymbol{a})+b$$

Регуляризующая функция $V(\boldsymbol{a}|\mu)$ и функция связи $q(y,z)$ выпуклы по $\boldsymbol{a}\in\mathbb{R}^n$ и $z\in\mathbb{R}$.

Вероятностные эвристики наблюдателя

1) Чем меньше параметрическая функция потерь $q\big(y,z(\boldsymbol{x},\boldsymbol{a},b)\big)$, тем более вероятно появление объекта с характеристиками $(\boldsymbol{x},y)$.

Предполагаемое параметрическое семейство совместных распределений $\varphi(\boldsymbol{x},y\,|\,\boldsymbol{a},b,c)\propto\exp\Big\{-cq\big(y,z(\boldsymbol{x},\boldsymbol{a},b)\big)\Big\}$.

2) Чем меньше регуляризующая функция $V(\boldsymbol{a}|\mu)$, тем более априори вероятно значение параметра $\boldsymbol{a}\in\mathbb{R}^n$.

Предполагаемое параметрическое семейство априорных распределений $\Psi(\boldsymbol{a}|\mu)\propto\exp\big\{-V(\boldsymbol{a}|\mu)\big\}$.

Выбор значений структурных параметров: **Максимизация полной функции правдоподобия (Evidence Function)** на обучающей совокупности

$$F(\boldsymbol{X},\boldsymbol{Y}|\mu,c)=\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n}\big[\Psi(\boldsymbol{a}|\mu)\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\big]d\boldsymbol{a}db\to\max(\mu,c),$$

$$\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)=\prod_{j=1}^{N}\varphi(\boldsymbol{x}_j,y_j\,|\,\boldsymbol{a},b,c).$$

Полная функция правдоподобия (Evidence Function)

Элементы вероятностной модели зависимости:

Априорная плотность распределения направляющего вектора и сдвига	$\Psi(\boldsymbol{a},b \mid \mu) \propto \exp\{-V(\boldsymbol{a} \mid \mu)\}$
Условная плотность распределения характеристик случайного объекта	$\varphi(\boldsymbol{x}, y \mid \boldsymbol{a},b,c) \propto \exp\{-cq\big(y, z(\boldsymbol{x},\boldsymbol{a},b)\big)\}$
Совместная условная плотность распределения случайной обучающей совокупности	$\Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a},b,c) = \prod_{j=1}^{N} \varphi(\boldsymbol{x}_j, y_j \mid \boldsymbol{a},b,c)$
Структурные параметры обобщенной линейной модели зависимости	– параметр μ регуляр. функции, – параметр c модели наблюдения.

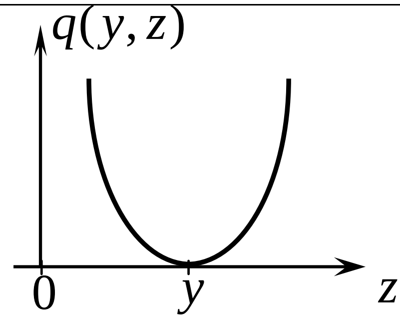
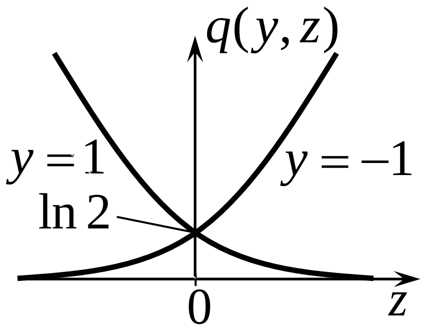
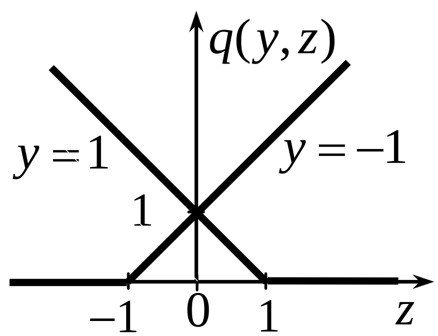
Маргинальная плотность распределения обучающей совокупности – непрерывная смесь условных распределений:

$F(\boldsymbol{X},\boldsymbol{Y} \mid \mu,c) = \int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}^n} \Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a},b,c)\Psi(\boldsymbol{a},b \mid \mu)d\boldsymbol{a}db$	Функция правдоподобия для (μ,c) (Marginal Likelihood, Evidence Function)

Оценки максимального правдоподобия для структурных параметров:
$(\hat{\mu},\hat{c}) = \operatorname{argmax} F(\boldsymbol{X},\boldsymbol{Y} \mid \mu,c)$

Вычислить этот интеграл легко только для простейших составляющих модели:	– функции связи $q(y,z)$, – регуляризующей функции $V(\boldsymbol{a} \mid \mu)$.

Нужен прямой алгоритм оптимизации, не требующий вычисления интеграла.

Некоторые функции связи	**Некоторые регуляризующие функции**
Числовая регрессия $y \in \mathbb{Y} = \mathbb{R}$, $q(y,z) = (y-z)^2$ 	Простейшая квадратичная регуляризация $V(\boldsymbol{a}) = \sum_{i=1}^{n} a_i^2$
Распознавание образов, логистическая регрессия $q(y,z) = \ln\left[1+\exp(-yz)\right]$ 	Квадратично-модульная регуляризация (Elastic Net) $V(\boldsymbol{a} \mid \mu) = \sum_{i=1}^{n} a_i^2 + \mu \sum_{i=1}^{n} \lvert a_i \rvert$
Распознавание образов, метод опорных векторов $q(y,z) = \max\left[0,\ 1-yz\right] = \begin{cases} 1-yz, & 1-yz > 0, \\ 0, & 1-yz \le 0. \end{cases}$ 	Другая квадратично-модульная регуляризация $V(\boldsymbol{a} \mid \mu) = \sum_{i=1}^{n} \begin{pmatrix} 2\mu \lvert a_i \rvert, & \lvert a_i \rvert \le \mu, \\ \mu^2 + a_i^2, & \lvert a_i \rvert > \mu. \end{pmatrix}$

Альтернативная запись функции правдоподобия

Два эквивалентных выражения для совместная плотность распределения наблюдений, направляющего вектора и сдвига:

$$\underbrace{H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{a},b\,|\,\mu,c)}_{\substack{\text{совместное}\\ \text{распределение}}}=\underbrace{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)}_{\substack{\text{условное}\\ \text{распределение}}}\;\underbrace{\Psi(\boldsymbol{a},b|\mu)}_{\substack{\text{априорное}\\ \text{распределение –}\\ \text{функция}\\ \text{правдоподобия}}}\;=\underbrace{F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)}_{\substack{\text{маргинальное}\\ \text{распределение}}}\underbrace{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}_{\substack{\text{апостериорное}\\ \text{распределение}}}$$

Отсюда прямое выражение для функции правдоподобия:

При любых значениях $(\boldsymbol{a},b)$

$$F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)=\frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}$$

Наиболее правдоподобные значения структурных параметров модели:

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)=\ln\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)+\ln\Psi(\boldsymbol{a},b|\mu)-\ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)\to\max(\mu,c)$$

Допустим, что мы можем вычислить апостериорную плотность для некоторых значений $(\hat{\mu}_k,\hat{c}_k)$, рассматриваемых как очередное приближение к точке максимума.

Теорема: Функция правдоподобия допускает представление:

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)=\int_{\mathbb{R}}\int_{\mathbb{R}^n}\left[\ln\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right]P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db+$$

$$\int_{\mathbb{R}}\int_{\mathbb{R}^n}\left[\ln\Psi(\boldsymbol{a},b|\mu)\right]P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db-\underbrace{\int_{\mathbb{R}}\int_{\mathbb{R}^n}\ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db}.$$

Фундаментальное неравенство:

$$\le\int_{\mathbb{R}}\int_{\mathbb{R}^n}\ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db$$

ЕМ-алгоритм максимизации функции правдоподобия

Допустим, что мы можем вычислить апостериорную плотность для некоторых значений $(\hat{\mu}_k, \hat{c}_k)$, рассматриваемых как очередное приближение к точке максимума.

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \int_{\mathbb{R}}\int_{\mathbb{R}^n}\left[\ln\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right]P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}\,db +$$

$$\int_{\mathbb{R}}\int_{\mathbb{R}^n}\left[\ln\Psi(\boldsymbol{a},b|\mu)\right]P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}\,db - \underbrace{\int_{\mathbb{R}}\int_{\mathbb{R}^n}\ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}\,db}.$$

Фундаментальное неравенство: $\geq \int_{\mathbb{R}}\int_{\mathbb{R}^n}\ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}\,db$

Отсюда следует правило пересчета $(\hat{\mu}_k, \hat{c}_k) \to (\hat{\mu}_{k+1}, \hat{c}_{k+1})$, гарантирующее неубывание функции правдоподобия:

$$\hat{\mu}_{k+1} = \arg\max_{\mu}\int_{\mathbb{R}}\int_{\mathbb{R}^n}\left[\ln\Psi(\boldsymbol{a},b|\mu)\right]P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}\,db,$$

$$\hat{c}_{k+1} = \underbrace{\arg\max_{\mu}}_{\text{maximization}}\ \underbrace{\int_{\mathbb{R}}\int_{\mathbb{R}^n}\left[\ln\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right]P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}\,db}_{\text{expectation}}.$$

Знаменитый **ЕМ-алгоритм** максимизации функций правдоподобия.

Теорема: $\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_{k+1},\hat{c}_{k+1}) \geq \ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_k,\hat{c}_k)$,

пока $\dfrac{\partial}{\partial\mu}\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_k,\hat{c}_k) \neq 0$ либо $\dfrac{\partial}{\partial c}\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_k,\hat{c}_k) \neq 0$.

Трудности реализации EM-алгоритма

EM-алгоритм:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a}, b | \mu) \right] P(\boldsymbol{a}, b \,|\, \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db,$$

$$\hat{c}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X}, \boldsymbol{Y} \,|\, \boldsymbol{a}, b, c) \right] P(\boldsymbol{a}, b \,|\, \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db.$$

Апостериорное распределение направляющего вектора и сдвига:

$$P(\boldsymbol{a}, b \,|\, \boldsymbol{X}, \boldsymbol{Y}, \mu, c) \propto \Psi(\boldsymbol{a}, b | \mu) \Phi(\boldsymbol{X}, \boldsymbol{Y} \,|\, \boldsymbol{a}, b, c)$$

Напомним, что здесь

$$\Psi(\boldsymbol{a}, b | \mu) \propto \exp\left\{ -V(\boldsymbol{a} | \mu) \right\}, \quad \Phi(\boldsymbol{X}, \boldsymbol{Y} \,|\, \boldsymbol{a}, b, c) \propto \exp\left\{ -c \sum_{j=1}^{N} q\Big(y_j, z(\boldsymbol{x}_j, \boldsymbol{a}, b) \Big) \right\}$$

Для многих регуляризующих функций $V(\boldsymbol{a}|\mu)$ и функций связи $q(y, z)$ интегрирование по такой апостериорной плотности затруднительно.

В следующем докладе Павел Левдик расскажет, как можно обойти эту трудность путем нормальной аппроксимации апостериорной плотности распределения:

$$P(\boldsymbol{a}, b \,|\, \boldsymbol{X}, \boldsymbol{Y}, \mu, c) \cong \mathcal{N}(\boldsymbol{a} \,|\, \hat{\boldsymbol{a}}_{\boldsymbol{y}, \boldsymbol{X}, \mu, c}, \boldsymbol{B}_{\boldsymbol{X}, c}) \mathcal{N}(b \,|\, \hat{b}_{\boldsymbol{y}, \boldsymbol{X}, \mu, c}(\boldsymbol{a}), \hat{\sigma}_N^2).$$