Прикладная статистика в машинном обучении
Семинар 9
Байесовский взгляд на подбор моделей

И. К. Козлов
(*Мехмат МГУ*)

2022

Настало время съесть лягушку!

Обсудим достаточные статистики экспоненциального класса!

Достаточные статистики

Достаточные статистики

Теорема факторизации Фишера

$T(x)$ — достаточная статистика $\Leftrightarrow$ существует разложение

$$f(x \mid \theta) = h(x)\, q(\theta, T(x)).$$

Докажем эту теорему в **дискретном случае**. В общем случае — см. Wikipedia:

https://en.wikipedia.org/wiki/Sufficient_statistic

Достаточные статистики

Доказательство теоремы факторизации.

($\Rightarrow$). Пусть $S(x)$ — достаточная, т.е. $\mathbb{P}(X \mid \mathrm{T} = t)$ не зависит от θ.

Обозначим $t = T(x)$. Тогда

$$f(x \mid \theta) = \mathbb{P}(X = x) = \underbrace{\mathbb{P}(X = x \mid \mathrm{T}(X) = t)}_{\text{не зависит от } \theta}\underbrace{\mathbb{P}(T = t)}_{\text{функция от } \theta \text{ и } \mathrm{T}}.$$

Это и есть нужное разложение:

$$f(x \mid \theta) = h(x)\, q(\theta, T(x)).$$

Достаточные статистики

($\Leftarrow$). (В дискретном случае). Пусть

$$f(x \mid \theta) = h(x)\, q(\theta, T(x)). \tag{1}$$

По определению условной вероятности

$$\mathbb{P}(X = x \mid \mathrm{T} = t) = \frac{\mathbb{P}(X = x, \mathrm{T} = t)}{\mathbb{P}(\mathrm{T} = t)} = \frac{\mathbb{P}(X = x)}{\mathbb{P}(\mathrm{T} = t)}.$$

Достаточные статистики

$(\Leftarrow)$. (В дискретном случае). Пусть

$$f(x \mid \theta) = h(x)\, q(\theta, T(x)). \tag{1}$$

По определению условной вероятности

$$\mathbb{P}(X = x \mid \mathrm{T} = t) = \frac{\mathbb{P}(X = x, \mathrm{T} = t)}{\mathbb{P}(\mathrm{T} = t)} = \frac{\mathbb{P}(X = x)}{\mathbb{P}(\mathrm{T} = t)}.$$

Числитель равен $f(x;\theta)$, а знаменатель — сумма вероятностей событий

$$\mathbb{P}(\mathrm{T} = t) = \sum_{x' : T(x') = t} f(x';\theta)$$

Подставляем (1):

$$\mathbb{P}(X = x \mid \mathrm{T} = t) = \frac{f(x;\theta)}{\sum_{x':T(x')=t} f(x';\theta)} = \frac{q(\theta, t) h(x)}{q(\theta, t) \sum_{x':T(x')=t} h(x')}$$

Достаточные статистики

$(\Leftarrow)$. (В дискретном случае). Пусть

$$f(x \mid \theta) = h(x)\, q(\theta, T(x)). \tag{1}$$

По определению условной вероятности

$$\mathbb{P}(X = x \mid \mathrm{T} = t) = \frac{\mathbb{P}(X = x, \mathrm{T} = t)}{\mathbb{P}(\mathrm{T} = t)} = \frac{\mathbb{P}(X = x)}{\mathbb{P}(\mathrm{T} = t)}.$$

Числитель равен $f(x;\theta)$, а знаменатель — сумма вероятностей событий

$$\mathbb{P}(\mathrm{T} = t) = \sum_{x' : T(x') = t} f(x';\theta)$$

Подставляем (1):

$$\mathbb{P}(X = x \mid \mathrm{T} = t) = \frac{f(x;\theta)}{\sum_{x':T(x')=t} f(x';\theta)} = \frac{q(\theta,t) h(x)}{q(\theta,t) \sum_{x':T(x')=t} h(x')}$$

$q(\theta, t)$ сокращается. Что и требовалось — $\mathbb{P}(X = x \mid \mathrm{T}(X) = t)$ не зависит от θ.

Экспоненциальное семейство распределений

Экспоненциальный класс

Семейство распределений, чья плотность может быть представлена в виде:

$$f(x|\theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right),$$

- $h(x) \geq 0$,
- $\theta^T u(x)$ означает $\theta_1 u_1(x) + \cdots + \theta_k u_k(x)$,
- Чтобы интеграл от плотности равнялся единице:

$$g(\theta) = \int h(x) \exp\left(\theta^T u(x)\right) dx.$$

Экспоненциальное семейство

Как мы помним, $u_j(x)$ — достаточные статистики:

$$f(x|\theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right).$$

Экспоненциальное семейство

Как мы помним, $u_j(x)$ — достаточные статистики:

$$f(x|\theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right).$$

Их моменты выражаются через производную $g(\theta)$:

- Матожидание

$$\mathbb{E}(u_j) = \frac{\partial \ln g(\theta)}{\partial \theta_j}.$$

- Матрица ковариации

$$\mathrm{Cov}(u_i, u_j) = \frac{\partial^2 \ln g(\theta)}{\partial \theta_i \partial \theta_j}.$$

Экспоненциальное семейство

Доказательство. Докажем для матожидания, для ковариации — аналогично.

- **Q**: Если $f(x|\theta) = \dfrac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right)$, то чему равна $g(\theta)$?

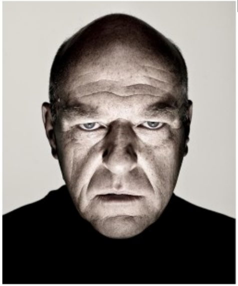
Доказательство. Докажем для матожидания, для ковариации — аналогично.

- **Q**: Если $f(x|\theta) = \dfrac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right)$, то чему равна $g(\theta)$?
- **A**: Это нормировочная константа:
$$g(\theta) = \int_{-\infty}^{\infty} h(x) \exp\left(\theta^T u(x)\right) dx.$$
- *Ну конечно, интеграл можно продифференцировать по параметру*:
$$\frac{\partial \ln g(\theta)}{\partial \theta_j} = \frac{1}{g(\theta)} \frac{\partial g(\theta)}{\partial \theta_j} = \int_{-\infty}^{\infty} u_j(x) \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right) dx = \mathbb{E}\left(u_j(x)\right).$$

Вычисления упрощаются

$$\mathbb{E}(u_j)$$

Интеграл считать



$$\frac{\partial \log g(\theta)}{\partial \theta_j}$$

Производную взять

Дифферецирование

Дифферециал

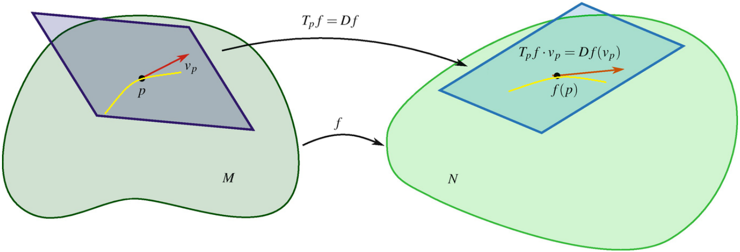
Дифферецирование

Наши мучения не заканчиваются)

Давайте вспомним — как дифференцировать функции.

Дифферецирование

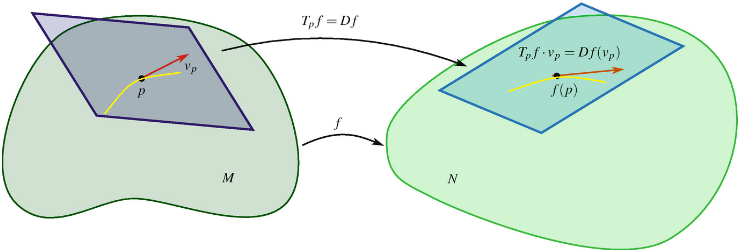
Пусть $f : M^n \to N^m$ — гладкое отображение. Тогда в каждой точке $p \in M^n$ определён дифференциал отображения
$$d_p f : T_p M^n \to T_{f(p)} N^m.$$



Дифферецирование

Пусть $f : M^n \to N^m$ — гладкое отображение. Тогда в каждой точке $p \in M^n$ определён дифференциал отображения

$$d_p f : T_p M^n \to T_{f(p)} N^m.$$



Жёлтая кривая $\gamma_M(t)$ на M переходит в жёлтую кривую $\gamma_N(t)$ на N.

Дифференциал $d_p f$ переводит вектор скорости $\dfrac{d\gamma_M}{dt}$ в вектор скорости $\dfrac{d\gamma_N}{dt}$.

Дифферецирование

Мы будем считать, что M^n *и* N^m — *области в* $\mathbb{R}^n$ *и* $\mathbb{R}^m$ соответственно.

Поэтому дифференциал будет линейным отображения

$$d_p f : \mathbb{R}^m \to \mathbb{R}^n$$

Будем писать $d_p f = df$ для краткости.

Как правильно вычислять в криволинейных координатах — учит дифференциальная геометрия.

Дифферецирование

Рассмотрим примеры:

- **Функция.** $f(x) : \mathbb{R} \to \mathbb{R}$. Какой вид имеет дифференциал?

Дифферецирование

Рассмотрим примеры:

- **Функция.** $f(x) : \mathbb{R} \to \mathbb{R}$. Какой вид имеет дифференциал?

$$df(x) = f'(x)dx$$

Дифферецирование

- **Функцию от вектора** $f(x^1, \dots, x^n) : \mathbb{R}^n \to \mathbb{R}$. Дифференциал?

[1] в евклидовых координатах

Дифферецирование

- **Функцию от вектора** $f(x^1, \ldots, x^n) : \mathbb{R}^n \to \mathbb{R}$. Дифференциал?

$$df(x) = \sum_{i=1}^{n} \frac{\partial f}{\partial x^i} dx^i.$$

Вектор частных производных[1] — градиент функции:

$$\nabla f = \left(\frac{\partial f}{\partial x^1}, \ldots, \frac{\partial f}{\partial x^n} \right).$$

Поэтому дифференциал — формальное скалярное произведение

$$df(x) = \langle \nabla f, dx \rangle$$

[1] в евклидовых координатах

Дифферецирование

- Далее — дифференциал **вектор-функции от вектора**

$$y^i = f^i(x^1, \dots, x^n), \qquad i = 1, \dots, m.$$

Дифферецирование

- Далее — дифференциал **вектор-функции от вектора**

$$y^i = f^i(x^1, \dots, x^n), \qquad i = 1, \dots, m.$$

Дифференциал

$$df(x) = J_f(x)dx,$$

где матрица Якоби

$$J_f = \begin{pmatrix} \frac{\partial f^1}{\partial x^1} & \cdots & \frac{\partial f^1}{\partial x^n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f^m}{\partial x^1} & \cdots & \frac{\partial f^m}{\partial x^n} \end{pmatrix}.$$

Дифферецирование

- **Функция от матрицы** $f(X)$, где X — $m \times n$ матрица.

  Q: Какой размер у матрицы дифференциала?

Дифферецирование

- **Функция от матрицы** $f(X)$, где X — $m \times n$ матрица.

  Q: Какой размер у матрицы дифференциала?

  A: $m \times n$ как у X. Пусть X^{ij} — компоненты X. Дифференциал:

$$df(X) = \sum_{i,j} \frac{\partial f}{\partial X^{ij}} dX^{ij} = \langle \nabla f, dX \rangle$$

Дифферецирование

Утверждение

Для любых $m \times n$ матриц A, B

$$\langle A, B \rangle = \operatorname{tr}(A^T B).$$

Дифферецирование

Утверждение

Для любых $m \times n$ матриц A, B

$$\langle A, B\rangle = \mathrm{tr}(A^T B).$$

Доказательство. Обозначим $A = (a_{ij})$, $B = (b_{ij})$ и $C = A^T B = (c_{ij})$. Тогда

$$c_{ij} = \sum_{k=1}^{m} a_{ki} b_{kj}.$$

Дифферецирование

Утверждение

Для любых $m \times n$ матриц A, B

$$\langle A, B \rangle = \mathrm{tr}(A^T B).$$

Доказательство. Обозначим $A = (a_{ij})$, $B = (b_{ij})$ и $C = A^T B = (c_{ij})$. Тогда

$$c_{ij} = \sum_{k=1}^{m} a_{ki} b_{kj}.$$

Поэтому

$$\mathrm{tr}(A^T B) = \sum_{i=1}^{n} \sum_{k=1}^{m} a_{ki} b_{ki} = \langle A, B \rangle.$$

Итого:

Вход \ Выход	Скаляр	Вектор	Матрица
Скаляр	$df(x) = f'(x)dx$ ($f'(x)$: скаляр; dx: скаляр)	–	–
Вектор	$df(x) = \langle \nabla f(x), dx \rangle$ ($\nabla f(x)$: вектор; dx: вектор)	$df(x) = J_f(x)dx$ ($J_f(x)$: матрица; dx: вектор)	–
Матрица	$df(X) = \langle \nabla f(X), dX \rangle$ ($\nabla f(X)$: матрица; dX: матрица)	–	–

Свойства диффереациала

Свойства дифференциала.

1. Это **линейный функционал**:

$$d(X+Y) = dX + dY, \qquad d(\alpha X) = \alpha d(X).$$

Если A, B — постоянные матрицы, то

$$d(AXB) = Ad(X)B.$$

Дифферецирование

2. Правило Лейбница

$$d(fg) = (df)g + f(dg)$$

С обычного произведения оно продолжается на матричные произведения:

$$d(XY) = (dX)Y + X(dY)$$

и скалярные произведения

$$d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle.$$

Дифферецирование

3. Производная сложной функции.

Дифференциал композиции отображений

$$M \xrightarrow{f} N \xrightarrow{g} P$$

будет композицией дифференциалов

$$d_x f(g(x)) = d_{f(x)} g \circ d_x g(x).$$

Дифферецирование

3. Производная сложной функции.

Дифференциал композиции отображений

$$M \xrightarrow{f} N \xrightarrow{g} P$$

будет композицией дифференциалов

$$d_x f(g(x)) = d_{f(x)} g \circ d_x g(x).$$

В координатах

$$\frac{\partial f(g(x))}{\partial x^i} = \sum_j \frac{\partial f}{\partial g^j} \frac{\partial g^j}{\partial x^i}.$$

Примеры вычисления

Дифферецирование

Q: Найти дифференциал

$$d\left(X^{-1}\right) = ?$$

Дифферецирование

Q: Найти дифференциал

$$d\left(X^{-1}\right) = ?$$

A: Дифференцируем тождество

$$X^{-1}X = E.$$

По правилу Лейбница

$$d\left(X^{-1}\right)X + X^{-1}dX = 0,$$

откуда

$$d\left(X^{-1}\right) = -X^{-1}d(X)X^{-1}.$$

Дифферецирование

Q: Найти градиент

$$\nabla \left(\operatorname{tr} AXBX^{-1}\right) = ?$$

Дифферецирование

Q: Найти градиент

$$\nabla\left(\operatorname{tr} AXBX^{-1}\right) = ?$$

A: Найдём его из свойства

$$\operatorname{tr}\left[(\nabla f)^T dX\right].$$

След — линейный оператор, поэтому

$$d(\operatorname{tr} X) = \operatorname{tr}(dX).$$

Далее используем правило Лейбница

$$d\left(\operatorname{tr} AXBX^{-1}\right) = \operatorname{tr}\left(Ad(X)BX^{-1}\right) + \operatorname{tr}\left(AXBd(X^{-1})\right).$$

Дифферецирование

Подставляя $d(X^{-1})$ получаем

$$d\left(\operatorname{tr} AXBX^{-1}\right) = \operatorname{tr}\left(Ad(X)BX^{-1}\right) - \operatorname{tr}\left(AXBX^{-1}d(X)X^{-1}\right).$$

Дифферецирование

Подставляя $d(X^{-1})$ получаем

$$d\left(\operatorname{tr} AXBX^{-1}\right) = \operatorname{tr}\left(Ad(X)BX^{-1}\right) - \operatorname{tr}\left(AXBX^{-1}d(X)X^{-1}\right).$$

Приводим выражение к виду

$$\operatorname{tr}\left[(\nabla f)^T dX\right],$$

используя свойства следа

$$\operatorname{tr} A = \operatorname{tr} A^T, \qquad \operatorname{tr} AB = \operatorname{tr} BA.$$

Дифферецирование

Подставляя $d(X^{-1})$ получаем

$$d\left(\operatorname{tr} AXBX^{-1}\right) = \operatorname{tr}\left(Ad(X)BX^{-1}\right) - \operatorname{tr}\left(AXBX^{-1}d(X)X^{-1}\right).$$

Приводим выражение к виду

$$\operatorname{tr}\left[(\nabla f)^T dX\right],$$

используя свойства следа

$$\operatorname{tr} A = \operatorname{tr} A^T, \qquad \operatorname{tr} AB = \operatorname{tr} BA.$$

Получаем

$$d\left(\operatorname{tr} AXBX^{-1}\right) = \operatorname{tr}\left[\left(A^T X^{-T} B^T A - X^{-T} B^T X^T A^T X^{-T}\right)^T dX\right]$$

Ответ:

$$\nabla f = A^T X^{-T} B^T A - X^{-T} B^T X^T A^T X^{-T}.$$

Дифферецирование

Q: Найти дифференциал

$$d \ln(\det X) = ?$$

Дифферецирование

Q: Найти дифференциал

$$d \ln (\det X) = ?$$

A: Производная сложной функции:

$$d \ln (\det X) = \frac{d (\det X)}{\det X}.$$

Далее вспомним пару фактов о матрицах.

Дифферецирование

Рассмотрим произвольную матрицу $A = (a_{ij})$.

Обозначим A_{ij} — алгебраические дополнения к a_{ij}.

Дифферецирование

Рассмотрим произвольную матрицу $A = (a_{ij})$.

Обозначим A_{ij} — алгебраические дополнения к a_{ij}.

- $\det A$ — многочлен от a_{ij};
- Разложения определителя по строке/столбцу:

$$\det A = \sum_{j=1}^{n} a_{ij} A_{ij} = \sum_{i=1}^{n} a_{ij} A_{ij}.$$

Дифферецирование

Рассмотрим произвольную матрицу $A = (a_{ij})$.

Обозначим A_{ij} — алгебраические дополнения к a_{ij}.

- $\det A$ — многочлен от a_{ij};
- Разложения определителя по строке/столбцу:

$$\det A = \sum_{j=1}^{n} a_{ij} A_{ij} = \sum_{i=1}^{n} a_{ij} A_{ij}.$$

Дифферецирование

1. частные производные определителя — алгебраические дополнения

$$\frac{\partial \det A}{\partial a_{ij}} = A_{ij}, \tag{2}$$

Дифферецирование

1. частные производные определителя — алгебраические дополнения

$$\frac{\partial \det A}{\partial a_{ij}} = A_{ij}, \tag{2}$$

2. элементы обратной матрицы — это алгебраические дополнения A_{ji}, делённые на определитель:

$$A^{-1} = \frac{1}{\det(A)}\,\mathrm{adj}(A), \qquad \text{где} \qquad \mathrm{adj}(A)_{ij} = A_{ji}.$$

Дифферецирование

Ответ:

$$d \ln(\det X) = \langle X^{-T}, dX \rangle.$$

Гессиан

Гессиан

Матрицу вторых производных называют гессианом и обозначают $\mathbf{H}_f$ или $\nabla^2 f$:

$$\nabla^2 f = \left(\frac{\partial^2 f}{\partial x_i \, \partial x_j} \right).$$

Формальное вычисление гессиана.

- Берём дифференциал
$$df(x) = \langle \nabla f(x), dx \rangle.$$
- Меняем $dx \to dx_1$.
- Ещё раз берём дифференциал
$$d\langle \nabla f(x), dx_1 \rangle = \langle d\left(\nabla f(x)\right), dx_1 \rangle = \langle \nabla^2 f(x) dx_2, dx_1 \rangle.$$

Гессиан

Q: Найти гессиан квадратичной функции

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + b^T x + c, \qquad x \in \mathbb{R}^n.$$

Гессиан

Q: Найти гессиан квадратичной функции

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + b^T x + c, \qquad x \in \mathbb{R}^n.$$

A: Находим дифференциал

$$df(x) = \langle Ax + b, dx\rangle.$$

И ещё раз берём дифференциал

$$d(Ax + b) = A.$$

Ответ: $\nabla^2 f = A$.

Памятка

Памятка

Правила преобразования
$dA = 0$
$d(\alpha X) = \alpha(dX)$
$d(AXB) = A(dX)B$
$d(X+Y) = dX + dY$
$d(X^T) = (dX)^T$
$d(XY) = (dX)Y + X(dY)$
$d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
$d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$

Таблица стандартных производных
$d\langle A, X\rangle = \langle A, dX\rangle$
$d\langle Ax, x\rangle = \langle (A + A^T)x, dx\rangle$
$d(\operatorname{tr} X) = \operatorname{tr}(dX)$
$d(\operatorname{Det}(X)) = \operatorname{Det}(X)\langle X^{-T}, dX\rangle$
$d(X^{-1}) = -X^{-1}(dX)X^{-1}$