

Прикладная статистика в машинном обучении

Лекция 5

Регрессионный анализ

И. К. Козлов
(Мехмат МГУ)

2022

Регрессия

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

Скатывание в посредственность

Регрессия — скатывание в посредственность.

Значение *regression* в английском



regression

noun [C or U]

UK /rɪ'greʃən/ US /rɪ'greʃən/

regression noun [C or U] (TO PREVIOUS STATE)



formal

a return to a previous and less advanced or worse state, condition, or way of behaving:

- A regression has occurred in the overall political situation.



Cambridge
Dictionary

Гальтон

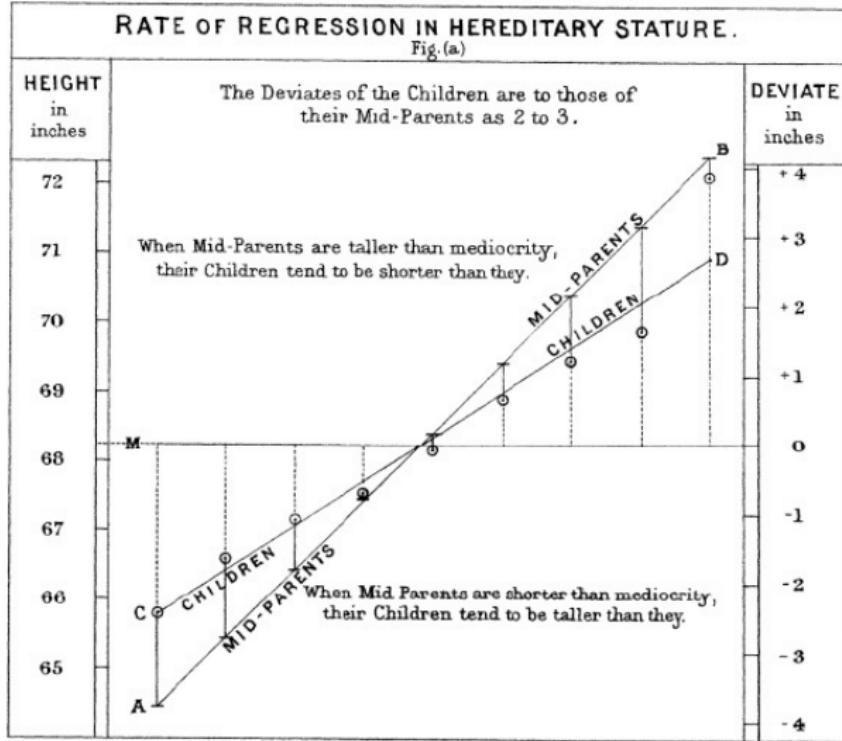
Ф. Гальтон, “Регрессия к посредственности в наследовании роста” (1885)

- X — рост родителя
- Y — рост ребёнка.
- Зависимость хорошо описывается уравнением

$$Y - \bar{Y} = \frac{2}{3} (X - \bar{X}).$$

RATE OF REGRESSION IN HEREDITARY STATURE.

Fig.(a)



Регрессия

- **Регрессия.** Даны пары $(x_1, Y_1), \dots, (x_n, Y_n)$, где

$$Y_i = r(x_i) + \varepsilon_i, \quad \mathbb{E}\varepsilon_i = 0.$$

Нужно восстановить функцию регрессии (**регрессор**) r .

Шум и гомоскедастичность

Будем считать, что шумы $\varepsilon_1, \dots, \varepsilon_n$

- имеют нулевое матожидание:

$$\mathbb{E}\varepsilon_i = 0,$$

- **гомоскедастичны**, т.е. имеют одинаковую дисперсию:

$$\mathbb{V}\varepsilon_i = \sigma^2$$

- некоррелированы:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j.$$

Линейная регрессия

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

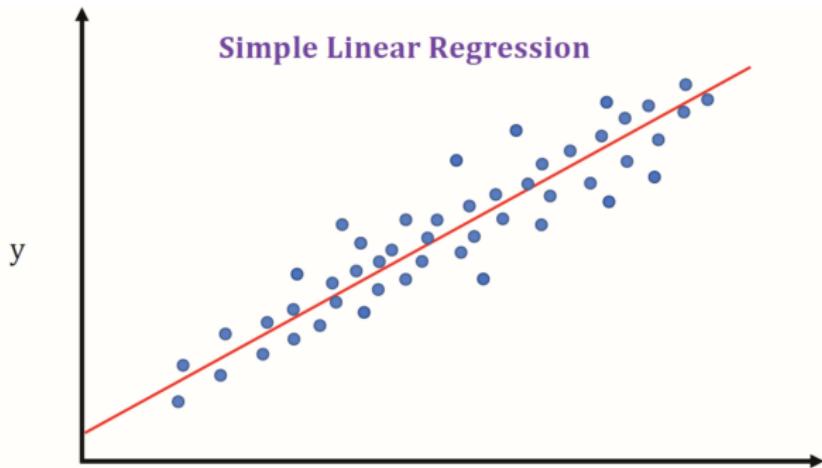
5 Feature Selection

6 А что нейронки?

Линейная регрессия

Рассмотрим простейшую задачу линейной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i, \quad i = 1, \dots, n,$$



Линейная регрессия

Вместе все эти уравнения можно записать в матричном виде

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Подчеркнём — чтобы учесть свободный член, в 1ом столбце матрицы \mathbf{X} стоят 1:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix},$$

Deja Vu

Deja Vu!



МНК = ОМП

На **Лекции 2** мы уже показали:

В предположении нормальности:

ОМП = оценка МНК

Сегодня будет больше Статистики.

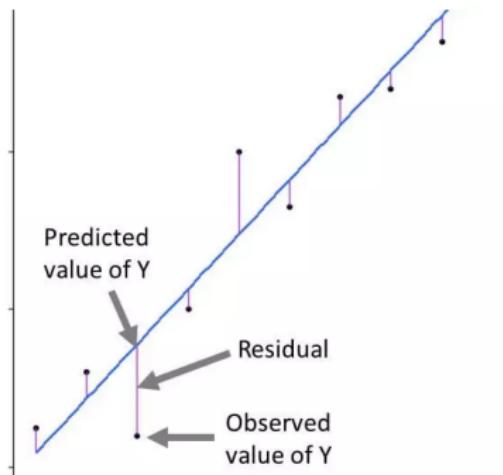
МНК-оценка

Оценка наименьших квадратов (МНК-оценка):

$$\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_m)^T$$

минимизирует остаточную сумму квадратов (Residual Sum of Squares):

$$RSS = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = \sum_{i=1}^n \hat{\varepsilon}_i^2.$$



МНК-оценка

Q: Формула для МНК оценки?

МНК-оценка

Q: Формула для МНК оценки?

A: МНК-оценка:

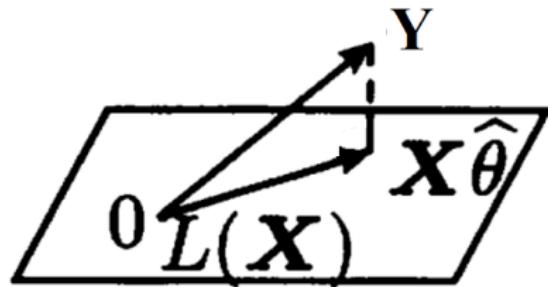
$$\hat{\beta} = X^+ Y.$$

Напомним формулу псевдообратной матрицы X^+ .

В разложении

$$Y = X\hat{\beta} + \hat{\varepsilon}$$

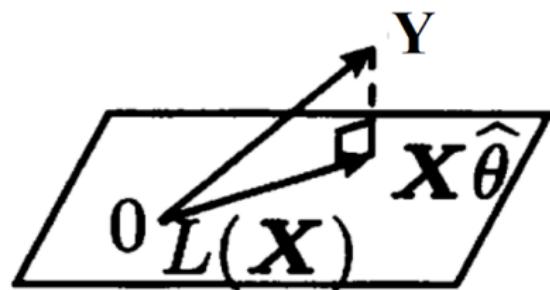
первый вектор лежит в $L(X)$ — подпространстве, порождённом X_1, \dots, X_n .



Q: Когда длина $\hat{\varepsilon}$ минимальна?

МНК-оценка

- $X\hat{\beta}$ — проекция Y на $L(X)$.
- Значит, $\hat{\varepsilon}$ ортогонален X_1, \dots, X_n .



МНК-оценка

Условие ортогональности

$$(X_i, \hat{\varepsilon}) = X_i^T \hat{\varepsilon} = 0$$

можно записать в виде

$$X^T \hat{\varepsilon} = X^T (Y - X\hat{\beta}) = 0$$

МНК-оценка

Условие ортогональности

$$(X_i, \hat{\varepsilon}) = X_i^T \hat{\varepsilon} = 0$$

можно записать в виде

$$X^T \hat{\varepsilon} = X^T (Y - X\hat{\beta}) = 0$$

МНК-оценка

Если $X^T X$ невырождена, то МНК-оценка:

$$\hat{\varepsilon} = (X^T X)^{-1} X^T Y.$$

МНК-оценка

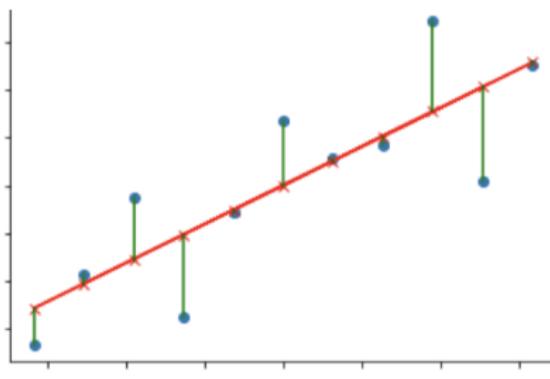
Далее будем везде считать, что $X^T X$ невырождена.

Следствие. МНК-оценка — это правильный ответ + линейный образ шума:

$$\begin{cases} \hat{\beta} = (X^T X)^{-1} X^T Y \\ Y = X\beta + \varepsilon \end{cases} \Rightarrow \hat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon.$$

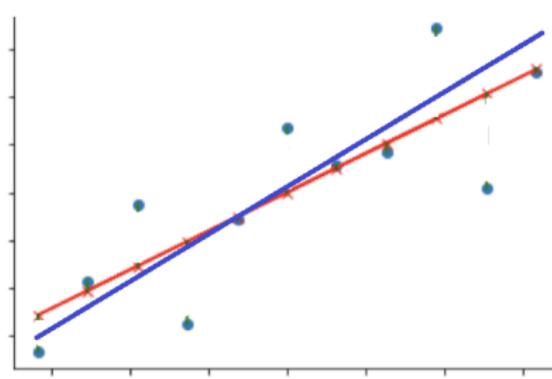
МНК-оценка

Изучим вопрос — насколько $\hat{\beta}$ является хорошей оценкой параметров β :



Truth: $Y = X\beta + \varepsilon$

↑
noise



Estimation: $Y = X\hat{\beta} + \hat{\varepsilon}$

↑
residual

Свойства МНК-оценок

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

Свойства МНК-оценок

Начнём со свойств МНК-оценок $\hat{\beta}$.

Это линейная алгебра — всё будет просто и замечательно.



Свойства МНК-оценок

Оценка $\hat{\beta}$ обладает всеми “хорошими” свойствами:

- Несмешённость $\mathbb{E}\hat{\beta} = \beta$;

Если $n\sigma^2(X^T X)^{-1} \rightarrow \Sigma$, то также

- Асимптотическая нормальность $\hat{\beta} \approx \mathcal{N}\left(\beta, \frac{\Sigma}{n}\right)$,
- Состоятельность $\hat{\beta} \xrightarrow{\mathbb{P}} \beta$;

Моменты

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

Начнём с вычисления моментов.

Теорема 1

Свойства МНК-оценки

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

- ① Оценка несмешённая:

$$\mathbb{E}\hat{\beta} = \beta.$$

- ② Матрица ковариации

$$\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

Моменты

Замечание. Формально эти матожидания условные:

$$\mathbb{E}(\hat{\beta} | X), \quad \text{Cov}(\hat{\beta} | X), \quad \dots$$

Для краткости мы просто пишем

$$\mathbb{E}(\hat{\beta}), \quad \text{Cov}(\hat{\beta}), \quad \dots$$

Замечание-2. Матрицу ковариации будем обозначать

$$\mathbb{V}(X) = \text{Cov}(X) = \text{Cov}(X, X).$$

Just Math It!

Доказательство — прямым вычислением.



just do it.

Моменты и линейное преобразование

Общие формулы. Пусть

- $\xi = \begin{pmatrix} \xi_1, \\ \dots \\ \xi_m \end{pmatrix}$ — произвольный случайный вектор, $\mathbb{E}\xi < \infty$, $\text{Cov}(\xi) < \infty$;
- A — это $k \times m$ матрица.

Q: $\mathbb{E}(A\xi) = ?$, $\text{Cov}(A\xi) = ?$

¹ Закон преобразования билинейной формы $Q' = C^T Q C$.

Моменты и линейное преобразование

Общие формулы. Пусть

- $\xi = \begin{pmatrix} \xi_1, \\ \dots \\ \xi_m \end{pmatrix}$ — произвольный случайный вектор, $\mathbb{E}\xi < \infty$, $\text{Cov}(\xi) < \infty$;
- A — это $k \times m$ матрица.

Q: $\mathbb{E}(A\xi) = ?$, $\text{Cov}(A\xi) = ?$

A: Матожидание линейно, ковариация — билинейна¹:

$$\mathbb{E}(A\xi) = A\mathbb{E}\xi, \quad \text{Cov}(A\xi) = A \text{Cov}(\xi) A^T.$$

¹ Закон преобразования билинейной формы $Q' = C^T Q C$.

Несмешённость

Несмешённость:

$$\mathbb{E}\hat{\beta} = \beta.$$

Доказательство:

$$\mathbb{E}\hat{\beta} = \mathbb{E}(\beta + X^+ \varepsilon) = \beta$$

Ковариация

Матрица ковариации:

$$\text{Cov}(\hat{\beta}) = \sigma^2(X^T X)^{-1}.$$

Доказательство:

$$\text{Cov}(\hat{\beta}) = \text{Cov}(\beta + X^+ \varepsilon) = X^+ \text{Cov}(\varepsilon)(X^+)^T.$$

Q: Чему равна $\text{Cov}(\varepsilon)$?

Ковариация

Матрица $\text{Cov}(\varepsilon)$ диагональна, т.к. $\mathbb{V}\varepsilon_i = \sigma^2$ и $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

$$\begin{aligned}\text{Cov}(\hat{\beta}) &= (X^T X)^{-1} X^T \sigma^2 \left((X^T X)^{-1} X^T \right)^T = \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.\end{aligned}$$

Теорема 1 доказана.

Асимптотика

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

Свойства МНК-оценок

Асимптотическая нормальность $\hat{\beta} \approx \mathcal{N}\left(\beta, \frac{\Sigma}{n}\right)$, если $n\sigma^2(X^T X)^{-1} \rightarrow \Sigma$.

Мы знаем моменты. Доказательство нормальности — см. Главу 21, §3, Теорема 3



М. Б. Лагутин,
Наглядная математическая статистика.

Свойства МНК-оценок

Вспоминаем **Лекцию 1**: Если $\text{Bias}(\beta_j) \rightarrow 0$ и $\mathbb{V}(\beta_j) \rightarrow 0$, то оценка состоятельна.

- Несмешённость $\Rightarrow \text{Bias} = 0$
- Асимптотическая нормальность $\hat{\beta} \approx \mathcal{N}\left(\beta, \frac{\Sigma}{n}\right) \Rightarrow$ ковариация $\frac{\Sigma}{n} \rightarrow 0$.

Следствие

МНК-оценка состоятельна:

$$\hat{\beta} \xrightarrow{\mathbb{P}} \beta.$$

Теорема Гаусса-Маркова

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

Вспомним ЛинАл

Продолжаем вспоминать **ЛинАл**.

- $\text{Cov}(X, Y)$ — это билинейная форма,
- $\mathbb{V}(X) = \text{Cov}(X) = \text{Cov}(X, X)$ — соответствующая квадратичная форма.

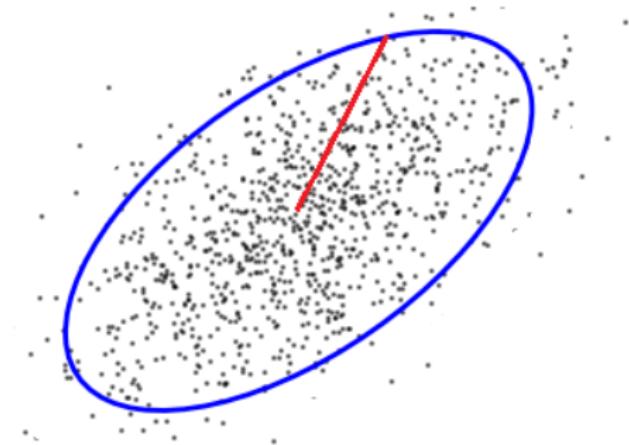


Дисперсия

Мы нашли матрицу ковариации $\text{Cov}(\hat{\beta})$.

Посмотрим на дисперсию в каждом направлении:

$$\mathbb{V}(c^T \hat{\beta}) = c^T \text{Cov}(\hat{\beta}) c.$$



Теорема Гаусса-Маркова

Итак,

$$\mathbb{E}\hat{\beta} = \beta, \quad \text{Cov}(\hat{\beta}) = \sigma^2(X^T X)^{-1}.$$

Q: Пусть $c \in \mathbb{R}^m$. Чему равны $\mathbb{E}(c^T \hat{\beta})$ и $\text{V}(c^T \hat{\beta})$?

Теорема Гаусса-Маркова

Итак,

$$\mathbb{E}\hat{\beta} = \beta, \quad \text{Cov}(\hat{\beta}) = \sigma^2(X^T X)^{-1}.$$

Q: Пусть $c \in \mathbb{R}^m$. Чему равны $\mathbb{E}(c^T \hat{\beta})$ и $\mathbb{V}(c^T \hat{\beta})$?

A: Матожидание линейно, дисперсия — квадратичная форма:

$$\mathbb{E}(c^T \hat{\beta}) = c^T \beta, \quad \mathbb{V}(c^T \hat{\beta}) = \sigma^2 c^T (X^T X)^{-1} c.$$

Теорема Гаусса-Маркова

OLS — ordinary least squares estimator.

BLUE — best linear unbiased estimator.

Gauss–Markov theorem:

OLS is **BLUE**

Теорема Гаусса-Маркова

Теорема Гаусса-Маркова

Среди **несмешённых линейных^a** оценок

$$\tilde{\beta}_j = c_{1j} Y_1 + \cdots + c_{nj} Y_n \quad \mathbb{E}(\tilde{\beta}) = \beta$$

МНК-оценка имеет наименьшую возможную дисперсию $\mathbb{V}(c^T \hat{\beta})$ для каждого $c \in \mathbb{R}^m$.

^aЛинейность по Y , коэффициенты c_{ij} могут зависеть от X .

Теорема Гаусса-Маркова

Теорема Гаусса-Маркова

Среди **несмешённых линейных^a** оценок

$$\tilde{\beta}_j = c_{1j} Y_1 + \cdots + c_{nj} Y_n \quad \mathbb{E}(\tilde{\beta}) = \beta$$

МНК-оценка имеет наименьшую возможную дисперсию $\mathbb{V}(c^T \hat{\beta})$ для каждого $c \in \mathbb{R}^m$.

^aЛинейность по Y , коэффициенты c_{ij} могут зависеть от X .

Замечание. Все условия важны.

Если шумы нормальны, то МНК - лучшая оценка в классе **всех несмешённых оценок**.

Перерыв

Перерыв

Доказательство теоремы Гаусса-Маркова

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

Теорема Гаусса-Маркова

Рассмотрим произвольную линейную оценку $\tilde{\beta} = CY$, где $C = X^+ + D$.

Шаг 1. Оценка несмешённая $\Leftrightarrow DX = 0$.

Теорема Гаусса-Маркова

Рассмотрим произвольную линейную оценку $\tilde{\beta} = CY$, где $C = X^+ + D$.

Шаг 1. Оценка несмешённая $\Leftrightarrow DX = 0$.

- По линейности

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E}[CY] = C\mathbb{E}[X\beta + \varepsilon] = CX\beta + C\mathbb{E}[\varepsilon].$$

Теорема Гаусса-Маркова

Рассмотрим произвольную линейную оценку $\tilde{\beta} = CY$, где $C = X^+ + D$.

Шаг 1. Оценка несмешённая $\Leftrightarrow DX = 0$.

- По линейности

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E}[CY] = C\mathbb{E}[X\beta + \varepsilon] = CX\beta + C\mathbb{E}[\varepsilon].$$

- По условию $\mathbb{E}[\varepsilon] = 0$, поэтому

$$\mathbb{E}[\tilde{\beta}] = ((X^T X)^{-1} X^T + D)X\beta = \beta + DX\beta.$$

- **Q:** Почему $DX = 0$?

Теорема Гаусса-Маркова

Рассмотрим произвольную линейную оценку $\tilde{\beta} = CY$, где $C = X^+ + D$.

Шаг 1. Оценка несмешённая $\Leftrightarrow DX = 0$.

- По линейности

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E}[CY] = C\mathbb{E}[X\beta + \varepsilon] = CX\beta + C\mathbb{E}[\varepsilon].$$

- По условию $\mathbb{E}[\varepsilon] = 0$, поэтому

$$\mathbb{E}[\tilde{\beta}] = ((X^T X)^{-1} X^T + D)X\beta = \beta + DX\beta.$$

- **Q:** Почему $DX = 0$?

Оценка несмешённая $\mathbb{E}[\tilde{\beta}] = \beta$ для любого $\beta \Leftrightarrow$ 2ое слагаемое равно нулю.

Теорема Гаусса-Маркова

Шаг 2. Дисперсия минимальна.

$$\begin{aligned}\mathbb{V}(\tilde{\beta}) &= \text{Var}(CY) = C \text{Var}(\varepsilon) C^T = \sigma^2 CC^T = \\ &= \sigma^2 \left((X^T X)^{-1} X^T + D \right) \left(X(X^T X)^{-1} + D^T \right)\end{aligned}$$

Теорема Гаусса-Маркова

Шаг 2. Дисперсия минимальна.

$$\begin{aligned}\mathbb{V}(\tilde{\beta}) &= \text{Var}(CY) = C \text{Var}(\varepsilon) C^T = \sigma^2 CC^T = \\ &= \sigma^2 \left((X^T X)^{-1} X^T + D \right) \left(X(X^T X)^{-1} + D^T \right)\end{aligned}$$

Честно раскрываем скобки

$$\mathbb{V}(\tilde{\beta}) = \sigma^2 (X^T X)^{-1} + \sigma^2 (X^T X)^{-1} (\textcolor{red}{DX})^T + \sigma^2 \textcolor{red}{DX} (X^T X)^{-1} + \sigma^2 DD^T.$$

Теорема Гаусса-Маркова

Шаг 2. Дисперсия минимальна.

$$\begin{aligned}\mathbb{V}(\tilde{\beta}) &= \text{Var}(CY) = C \text{Var}(\varepsilon) C^T = \sigma^2 CC^T = \\ &= \sigma^2 \left((X^T X)^{-1} X^T + D \right) \left(X(X^T X)^{-1} + D^T \right)\end{aligned}$$

Честно раскрываем скобки

$$\mathbb{V}(\tilde{\beta}) = \sigma^2 (X^T X)^{-1} + \sigma^2 (X^T X)^{-1} (\textcolor{red}{DX})^T + \sigma^2 \textcolor{red}{DX} (X^T X)^{-1} + \sigma^2 DD^T.$$

Согласно Шагу 1 $\textcolor{red}{DX} = 0$.

Q: Чему равно 1ое слагаемое?

Теорема Гаусса-Маркова

Мы знаем, что $\sigma^2(X^T X)^{-1} = \mathbb{V}(\hat{\beta})$. Поэтому

$$\mathbb{V}(\tilde{\beta}) = \mathbb{V}(\widehat{\beta}) + \sigma^2 D D^T.$$

Теорема Гаусса-Маркова

Мы знаем, что $\sigma^2(X^T X)^{-1} = \mathbb{V}(\hat{\beta})$. Поэтому

$$\mathbb{V}(\tilde{\beta}) = \mathbb{V}(\hat{\beta}) + \sigma^2 D D^T.$$

Получаем требуемое:

$$\mathbb{V}(c^T \tilde{\beta}) - \mathbb{V}(c^T \hat{\beta}) = \sigma^2 c^T D D^T c \geq 0.$$

Теорема Гаусса-Маркова доказана.

Feature Selection

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

Насущный вопрос

Насущный вопрос — пилить ли фичу?



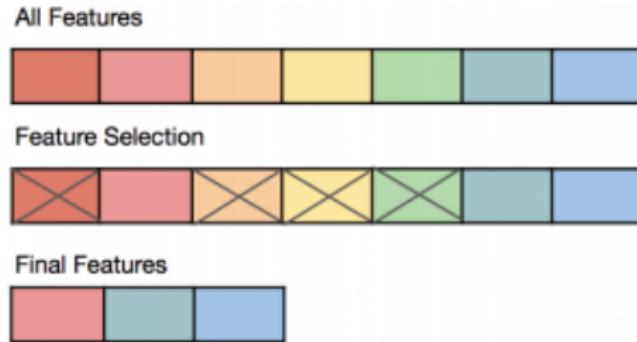
-..Пилите, Шура, пилите!..

Рис.: Или всё-таки не пилить?

Feature Selection

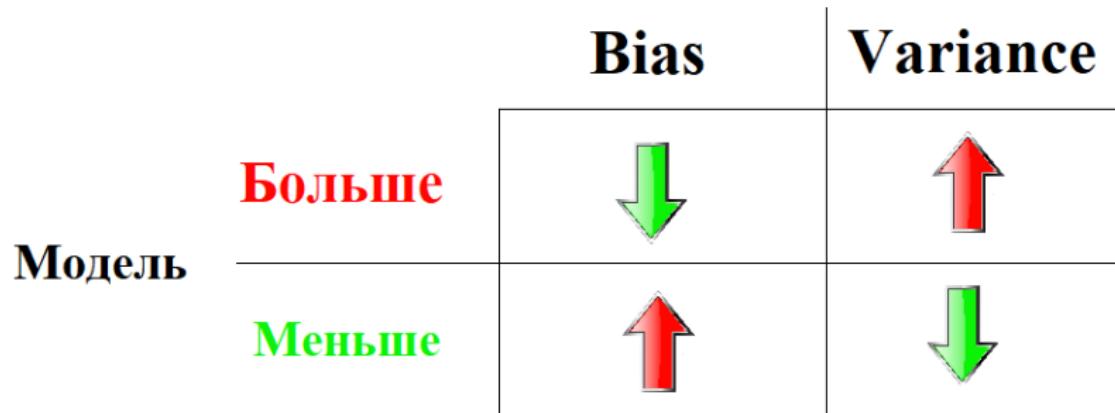
Поговорим об отборе признаков (Feature Selection) для моделей.

Q: Зачем отбирать фичи/признаки для модели?



Bias-Variance trade-off

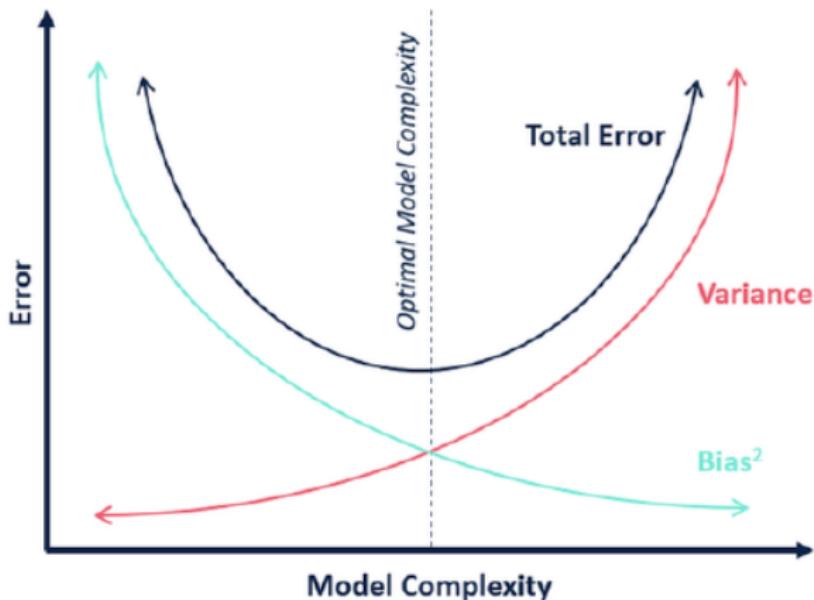
Для линейных моделей есть Bias-Variance trade-off.



Bias-Variance trade-off

Простые модели
недообучаются
(underfitting).

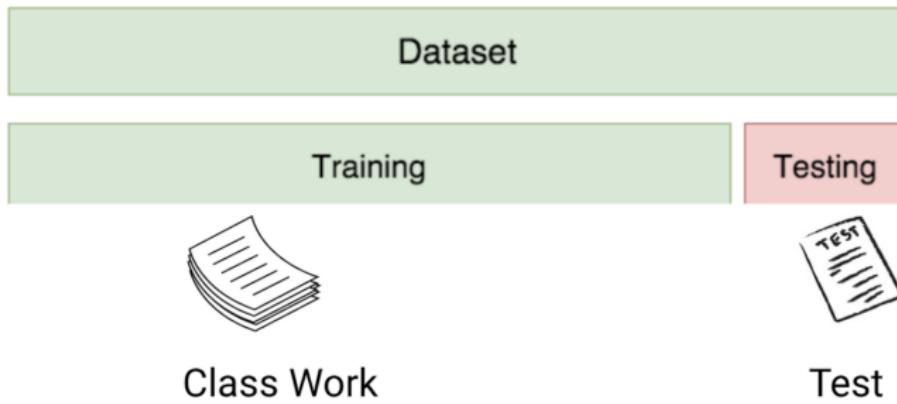
Сложные модели
переобучаются
(overfitting).



Сдача экзамена

Приведём неформальное² рассуждение.

Представим обучение ML-модели как “сдачу экзамена”.



²Нестрогое и неверное!

Сдача экзамена

Сдавали экзамен 3 модели:

Простая Модель



Ничего не выучила

Train 5%
Test 1%

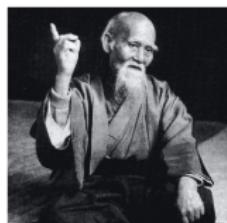
Сложная модель



Запомнила все ответы

Train 100%
Test 0%

Оптимальная модель



Познала суть

Train 98%
Test 92%

Lack of fit + complexity penalty

Q: Как оценивать модели?

A: Кажутся разумными критерии вида

Ошибка модели + Сложность модели

Статистики

Пусть

- Всего n признаков в полной модели.
- Из них мы отобрали k признаков.
- L — правдоподобие для меньшей модели.
- \hat{Y}_i — предсказания меньшей модели.

AIC, BIC, Mallow Cp

Есть несколько известных статистик (все минимизируются):

- ① Информационный критерий Акаике (1974)

$$AIC = 2k - 2 \ln(L);$$

AIC, BIC, Mallow Cp

Есть несколько известных статистик (все минимизируются):

- ① Информационный критерий Акаике (1974)

$$AIC = 2k - 2 \ln(L);$$

- ② Байесовский информационный критерий (1978)

$$BIC = k \ln(n) - 2 \ln(L);$$

AIC, BIC, Mallow Cp

Есть несколько известных статистик (все минимизируются):

- ① Информационный критерий Акаике (1974)

$$AIC = 2k - 2 \ln(L);$$

- ② Байесовский информационный критерий (1978)

$$BIC = k \ln(n) - 2 \ln(L);$$

- ③ Статистика C_P Mallow (1973)

$$\hat{R} = \hat{R}_{\text{tr}} + 2k\hat{\sigma}^2,$$

- $\hat{R}_{\text{tr}} = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$ — ошибка на обучающей выборке
- $\hat{\sigma}^2$ — оценка дисперсии ошибки по полной модели (со всеми признаками).

Meh

На практике особого толку от этих доисторических статистик **НЕТ**.



No silver bullet

Универсального алгоритма для отбора признаков **НЕТ**.



Отбор признаков регрессией

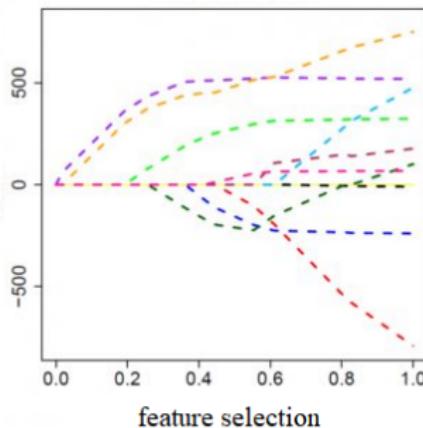
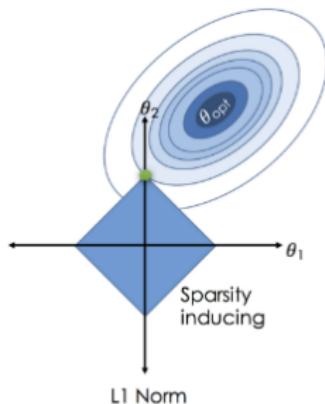
Q: Мы уже сталкивались с моделью, которая может отбирать признаки?

Отбор признаков регрессией

Q: Мы уже сталкивались с моделью, которая может отбирать признаки?

LASSO

(Регрессия + L1-регуляризация)



На практике — последовательно добавляем-удаляем признаки ([Stepwise regression](#)).

Q: Как сравнить модели?

Cross-validation

На практике — последовательно добавляем-удаляем признаки ([Stepwise regression](#)).

Q: Как сравнить модели?

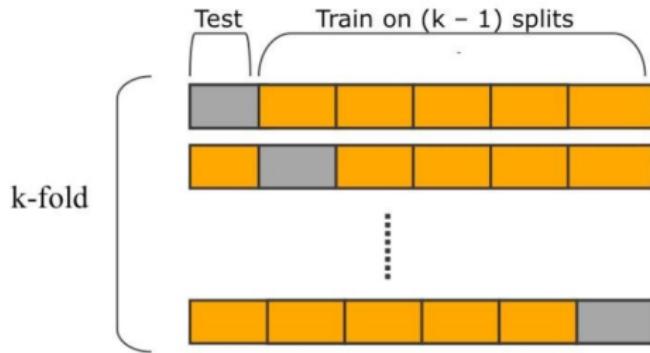
A: Сравнить средние ошибки скользящего контроля ([Кросс-валидация](#)).

Cross-validation

На практике — последовательно добавляем-удаляем признаки ([Stepwise regression](#)).

Q: Как сравнить модели?

A: Сравнить средние ошибки скользящего контроля ([Кросс-валидация](#)).



А что нейронки?

1 Регрессия

2 Линейная регрессия

3 Свойства МНК-оценок

- Моменты
- Асимптотика

4 Теорема Гаусса-Маркова

- Доказательство теоремы Гаусса-Маркова

5 Feature Selection

6 А что нейронки?

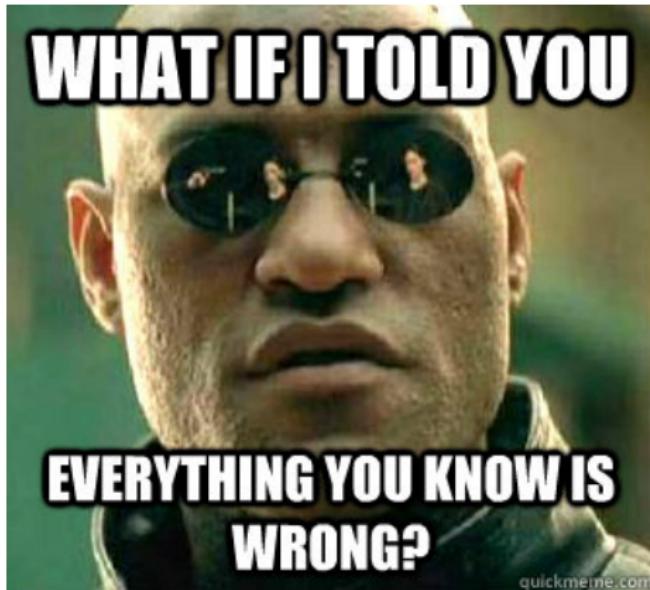
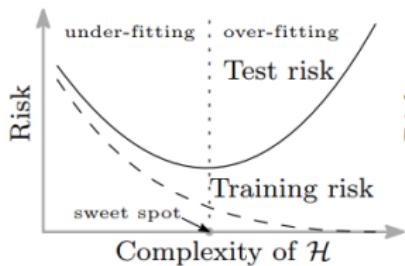


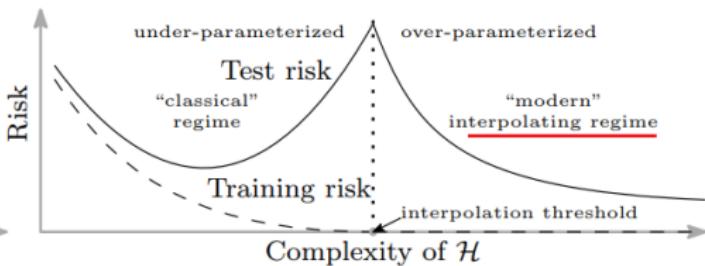
Рис.: А потом пришли нейронки

Double Decent

Был открыт эффект [Двойного Спуска](#) (Double Decent), противоречащий Bias-Variance Trade-off.



(a) U-shaped “bias-variance” risk curve



(b) “double descent” risk curve



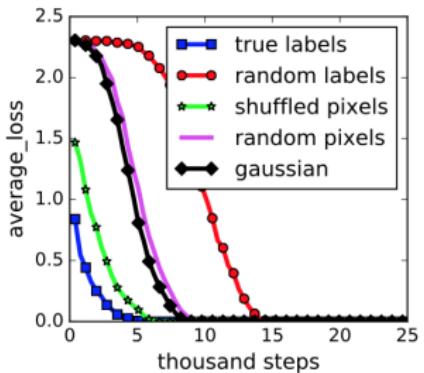
M. Belkin, D. Hsu, S. Ma, and S. Mandal. (2012)

Reconciling modern machine learning practice and the bias-variance trade-off.

DNN can fit randomly labeled data

Чудеса продолжаются!

Нейронки способны выучить всё, что угодно,
даже **randomные метки** (только медленней).

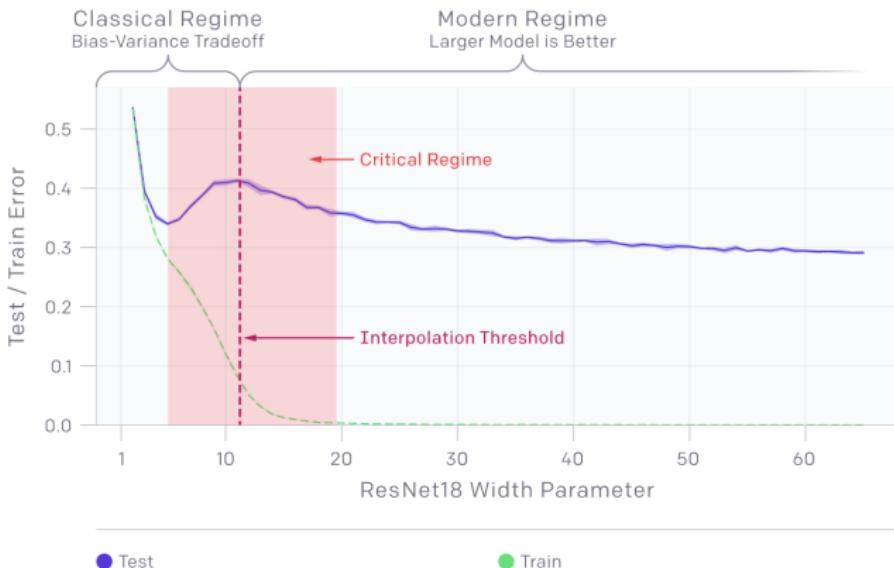


C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals (2016)
Understanding deep learning requires rethinking generalization

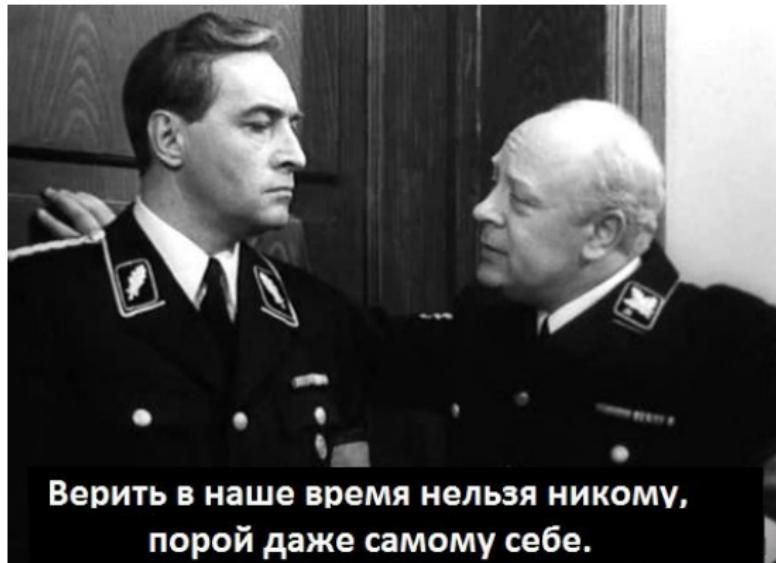
CNN, ResNet, Трансформеры — не переобучаются!

Подробнее в популярном блоге OpenAI:

<https://openai.com/blog/deep-double-descent/>



Главное, что мы узнали сегодня:



Верить в наше время нельзя никому,
порой даже самому себе.

← To Be Continued |||