# Занятие 3
# Линейные методы регрессии.

Елена Кантонистова

ekantonistova@hse.ru

# ПЛАН ЛЕКЦИИ

- Линейная регрессия (напоминание)
- Почему MSE? Вероятностное объяснение
- Точное решение (OLS или метод наименьших квадратов)
- Особенности применения линейной регрессии

# ЛИНЕЙНАЯ РЕГРЕССИЯ (НАПОМИНАНИЕ)

# ЛИНЕЙНАЯ РЕГРЕССИЯ

Пример (напоминание):

Предположим, что мы хотим предсказать *стоимость дома* $y$ по его *площади* $(x_1)$ и *количеству комнат* $(x_2)$.

Линейная модель для предсказания стоимости:

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2},$$

где $w_0, w_1, w_2$ -

параметры модели (*веса*).

# ЛИНЕЙНАЯ РЕГРЕССИЯ

Пример (напоминание):

Предположим, что мы хотим предсказать *стоимость дома* $y$

по его *площади* $(x_1)$ и *количеству комнат* $(x_2)$.

Линейная модель для предсказания стоимости:

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2},$$

где $w_0, w_1, w_2$ -

параметры модели (*веса*).

Общий вид (линейная регрессия):

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n},$$

где $x_1, \ldots, x_n$ - признаки объекта $x$.

# ЛИНЕЙНАЯ РЕГРЕССИЯ

Линейная регрессия:

$$a(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

# ЛИНЕЙНАЯ РЕГРЕССИЯ

Линейная регрессия:

$$a(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

- сокращенная запись:

$$a(x) = w_0 + \sum_{j=1}^{n} w_j x_j$$

# ЛИНЕЙНАЯ РЕГРЕССИЯ

Линейная регрессия:

$$a(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

- сокращенная запись:

$$a(x) = w_0 + \sum_{j=1}^{n} w_j x_j$$

- запись через скалярное произведение (с добавлением признака $x_0 = 1$):

$$a(x) = w_0 \cdot 1 + \sum_{j=1}^{n} w_j x_j = \sum_{j=0}^{n} w_j x_j = (w, x)$$

# ЛИНЕЙНАЯ РЕГРЕССИЯ

Линейная регрессия:

$$a(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

- сокращенная запись:

$$a(x) = w_0 + \sum_{j=1}^{n} w_j x_j$$

- запись через скалярное произведение (с добавлением признака $x_0 = 1$):

$$a(x) = w_0 \cdot 1 + \sum_{j=1}^{n} w_j x_J = \sum_{j=0}^{n} w_j x_j = (w, x) \leftrightarrow a(x) = (w, x)$$

# ЛИНЕЙНАЯ РЕГРЕССИЯ

Линейная регрессия:

$$a(x) = w_0 + \sum_{j=1}^{n} w_j x_j = (w, x)$$

**Обучение линейной регрессии - минимизация среднеквадратичной ошибки:**

$$\mathrm{Q}(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 = \frac{1}{l}\sum_{i=1}^{l}((w, x_i) - y_i)^2 \to \min_w$$

(здесь $l$ – количество объектов)

# ПОЧЕМУ MSE?

# ВЕРОЯТНОСТНАЯ ПОСТАНОВКА

- Даже если целевая переменная линейно зависит от признаков, то идеальной модели не существует, то есть реальные ответы будут отличаться от предсказаний, поэтому мы пишем

$$y \approx (w, x)$$

- Неидеальность прогноза можно объяснить неполнотой данных, или же шумом в данных. Тогда формула переписывается со знаком “=”:

$$y = (w, x) + \varepsilon,$$

где $\varepsilon$ – шум в данных.

# ВЕРОЯТНОСТНАЯ ПОСТАНОВКА

$$y = (w, x) + \varepsilon$$

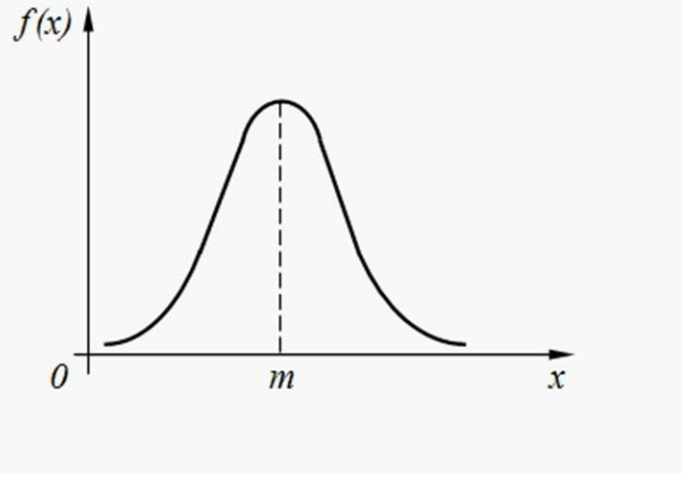
- Шум в данных обычно имеет некоторое распределение. В большинстве реальных задач считается, что

$$\varepsilon \sim N(0, \sigma^2).$$

- Отсюда получаем, что

$y \sim N\big((w, x), \sigma^2\big)$.

*График плотности нормального распределения*



# ВЕРОЯТНОСТНАЯ ПОСТАНОВКА

$$y \sim N\big((w,x),\sigma^2\big)$$

Это означает, что вероятность наблюдать $y$ при данных значениях $x$ равна

$$p(y|x,w) \sim N((w,x),\sigma^2)$$

**Мы хотим подобрать оптимальные веса. Что это значит?**

Мы хотим подобрать такой вектор $w$, что вероятность наблюдать некоторое значение $y$ при наблюдаемых $x$ максимальна.

# МЕТОД МАКСИМУМА ПРАВДОПОДОБИЯ (ММП)

**Мы хотим подобрать оптимальные веса. Что это значит?**

Мы хотим подобрать такой вектор $w$, что вероятность наблюдать некоторое значение $y$ при наблюдаемых $x$ максимальна.

Запишем это желание сразу для всех объектов выборки (в предположении, что объекты независимы):

$$p(\boldsymbol{y}|X,w) = p(y_1|x_1,w) \cdot p(y_2|x_2,w) \cdot \ldots \cdot p(y_i|x_i,w) \cdot \ldots \to \max_w$$

Величина $p(y|X,w)$ называется ***функцией правдоподобия (или правдоподобием) выборки***.

# ММП ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Модель данных с некоррелированным гауссовским шумом:

$$y_i = (w, x_i) + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, l$$

# ММП ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Модель данных с некоррелированным гауссовским шумом:

$$y_i = (w, x_i) + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, l$$

Тогда $y_i \sim N((w, x_i), \sigma^2), i = 1, \dots, l$

# ММП ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Модель данных с некоррелированным гауссовским шумом:

$$y_i = (w, x_i) + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, l$$

Тогда $y_i \sim N((w, x_i), \sigma^2), i = 1, \dots, l$

Метод максимума правдоподобия (ММП):

$$L(y_1, \dots, y_l | w) = \prod_{i=1}^{l} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - (w, x_i)\right)^2\right) \to \max_w$$

# ММП ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Модель данных с некоррелированным гауссовским шумом:

$$y_i = (w, x_i) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, l$$

Тогда $y_i \sim N((w, x_i), \sigma^2), i = 1, \dots, l$

Метод максимума правдоподобия (ММП):

$$L(y_1, \dots, y_l | w) = \prod_{i=1}^{l} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - (w, x_i)\right)^2\right) \to \max_w$$

$$-ln\ L(y_1, \dots, y_l | w) = const + \frac{1}{2\sigma^2}\sum_{i=1}^{l}\left(y_i - (w, x_i)\right)^2 \to \min_w$$

В данном случае ММП совпадает с МНК.

# МАТРИЧНОЕ ДИФФЕРЕНЦИРОВАНИЕ И АНАЛИТИЧЕСКОЕ РЕШЕНИЕ ЗАДАЧИ МНК

# МАТРИЧНОЕ ДИФФЕРЕНЦИРОВАНИЕ

Напомним, как выглядит производная функции-скаляра по вектору (1) и по матрице (2):

1) $$\nabla_{\mathbf{x}} f = \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n.$$

2) $$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \frac{\partial f}{\partial X_{m2}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

**Задача 1.** *Пусть $a \in \mathbb{R}^n$ — вектор параметров, а $x \in \mathbb{R}^n$ — вектор переменных. Необходимо найти производную их скалярного произведения по вектору переменных $\nabla_x a^T x$.*

**Задача 1.** *Пусть $a \in \mathbb{R}^n$ — вектор параметров, а $x \in \mathbb{R}^n$ — вектор переменных. Необходимо найти производную их скалярного произведения по вектору переменных $\nabla_x a^T x$.*

**Решение.**

$$\frac{\partial}{\partial x_i} a^T x = \frac{\partial}{\partial x_i} \sum_j a_j x_j = a_i,$$

поэтому $\nabla_x a^T x = a$.

Заметим, что $a^T x$ — это число, поэтому $a^T x = x^T a$, следовательно,

$$\nabla_x x^T a = a.$$

**Задача 2.** *Пусть теперь* $A \in \mathbb{R}^{n \times n}$. *Необходимо найти* $\nabla_x x^T A x$.

**Задача 2.** *Пусть теперь $A \in \mathbb{R}^{n \times n}$. Необходимо найти $\nabla_x x^T Ax$.*

**Решение.**

$$\frac{\partial}{\partial x_i} x^T Ax = \frac{\partial}{\partial x_i} \sum_j x_j (Ax)_j = \frac{\partial}{\partial x_i} \sum_j x_j \left( \sum_k a_{jk} x_k \right) = \frac{\partial}{\partial x_i} \sum_{j,k} a_{jk} x_j x_k =$$

$$= \sum_{j \neq i} a_{ji} x_j + \sum_{k \neq i} a_{ik} x_k + 2a_{ii} x_i = \sum_j a_{ji} x_j + \sum_k a_{ik} x_k = \sum_j (a_{ji} + a_{ij}) x_j.$$

Поэтому $\nabla_x x^T Ax = (A + A^T)x$.

$\nabla_x x^T a = a.$

# АНАЛИТИЧЕСКОЕ РЕШЕНИЕ ЗАДАЧИ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ (МНК)

Задача обучения линейной регрессии (в матричной форме):

$$\frac{1}{\ell}\|Xw - y\|^2 \to \min_w$$

Точное (аналитическое) решение [с выводом на доске]:

$$w = (X^T X)^{-1} X^T y$$

# НЕДОСТАТКИ АНАЛИТИЧЕСКОЙ ФОРМУЛЫ

- Обращение матрицы – сложная операция ($O(N^3)$ от числа признаков)
- Матрица $X^T X$ может быть вырожденной или плохо обусловленной
- Если заменить среднеквадратичный функционал ошибки на другой, то скорее всего не найдем аналитическое решение

# ОСОБЕННОСТИ ПРИМЕНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

# О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Пример:

Предположим, что мы хотим предсказать *стоимость дома* $y$ по его *площади* ($x_1$) и *количеству комнат* ($x_2$), *району* ($x_3$) и *удаленности от МКАД* ($x_4$).

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4.}$$

# О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Пример:

Предположим, что мы хотим предсказать *стоимость дома* $y$ по его *площади* ($x_1$) и *количеству комнат* ($x_2$), *району* ($x_3$) и *удаленности от МКАД* ($x_4$).

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4.}$$

Проблема №1: район ($x_3$) – это не число, а название района. Например, Мамыри, Дудкино, Барвиха… Что с этим делать?

# О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Пример:

Предположим, что мы хотим предсказать *стоимость дома* $y$ по его *площади* ($x_1$) и *количеству комнат* ($x_2$), *району* ($x_3$) и *удаленности от МКАД* ($x_4$).

$$a(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4.$$

Проблема №1: район ($x_3$) – это не число, а название района. Например, Мамыри, Дудкино, Барвиха… Что с этим делать?

Решение – one-hot encoding (OHE): создаем новые числовые столбцы, каждый из которых является индикатором района.

# ONE-HOT ENCODING

| Район |
|---|
| Дудкино |
| Барвиха |
| Мамыри |
| … |
| Барвиха |

| Мамыри | Дудкино | Барвиха |
|---|---|---|
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| … | … | … |
| 0 | 0 | 1 |

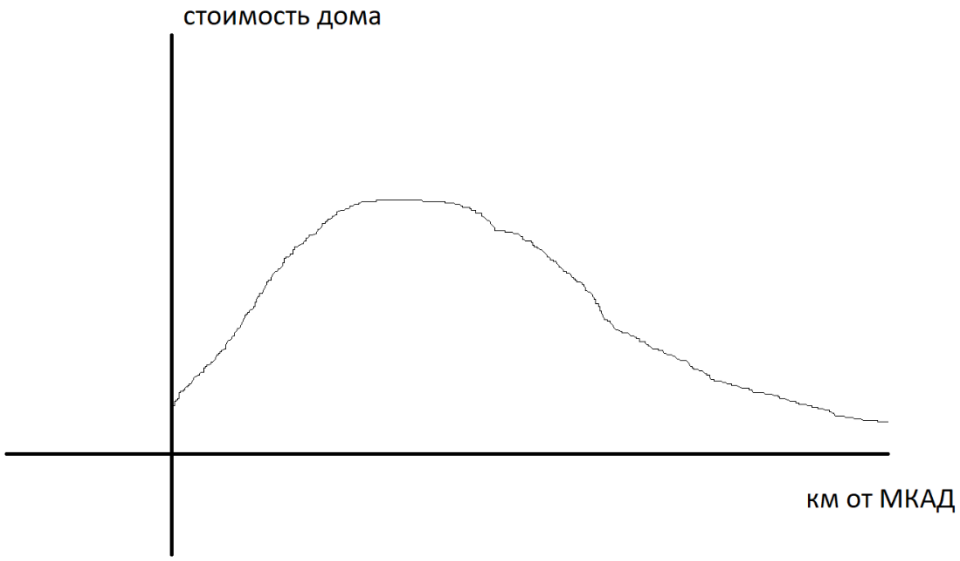
$$\boldsymbol{a(x)} =$$
$$= \boldsymbol{w_0} + \boldsymbol{w_1 x_1} + \boldsymbol{w_2 x_2} + \boldsymbol{w_{31}} \boldsymbol{x}_{\text{Мамыри}} + \boldsymbol{w_{32}} \boldsymbol{x}_{\text{Дудкино}} + \boldsymbol{w_{33}} \boldsymbol{x}_{\text{Барвиха}} + \boldsymbol{w_4}\, \boldsymbol{x_4}.$$

# О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Пример:

Предположим, что мы хотим предсказать *стоимость дома* $y$ по его *площади* $(x_1)$ и *количеству комнат* $(x_2)$, *району* $(x_3)$ и *удаленности от МКАД* $(x_4)$.

Проблема №2: удаленность от МКАД ($x_4$) не монотонно влияет на стоимость дома.



# О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Проблема №2: удаленность от МКАД ($x_4$) не монотонно влияет на стоимость дома.

Решение – бинаризация (разбиение на бины).

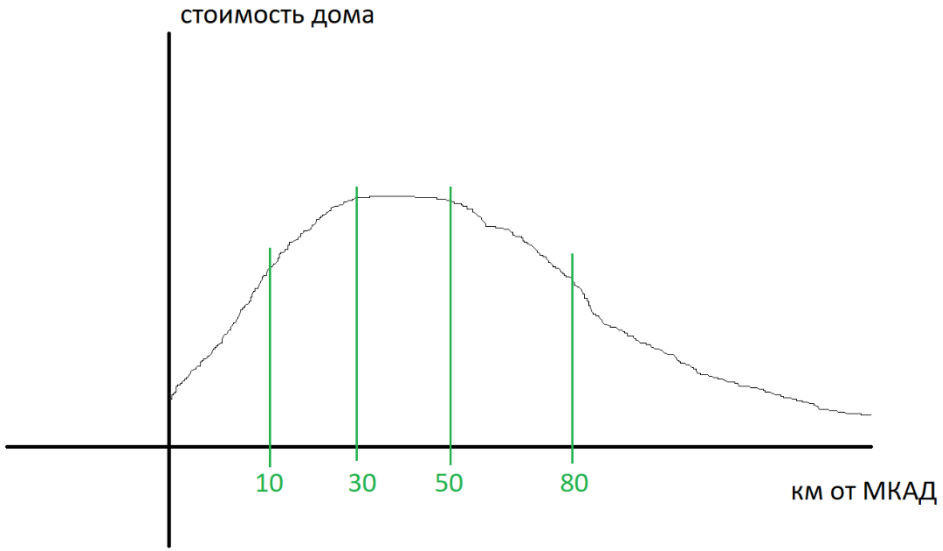
Новые признаки:

- $x_{[0;10)}$ - равен 1, если

дом находится в пределах

10 км от МКАД, и 0 иначе

- $x_{[10;30)}$ - равен 1, если

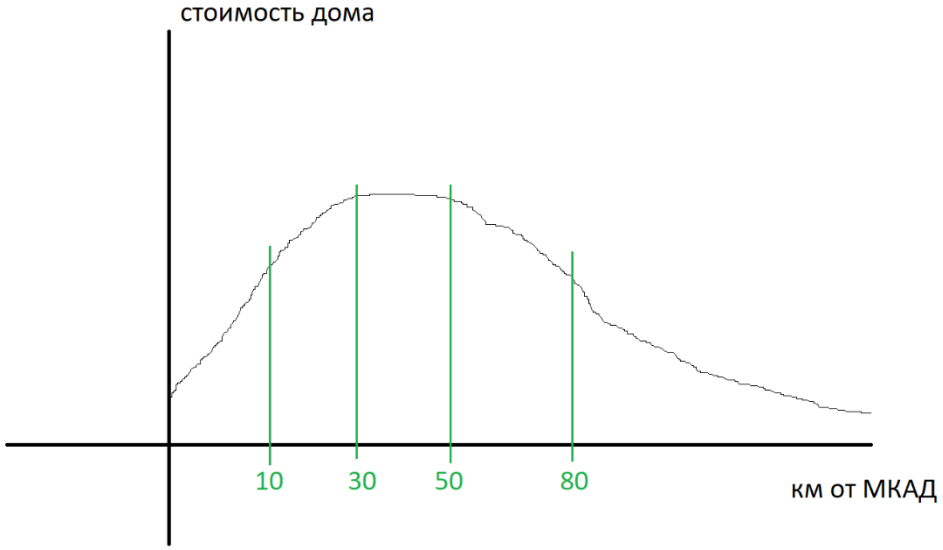



дом находится в пределах от 10 км до 30 км МКАД, и 0 иначе. И т.д.

# О ПРИМЕНИМОСТИ ЛИНЕЙНОЙ МОДЕЛИ

Проблема №2: удаленность от МКАД ($x_4$) не монотонно влияет на стоимость дома.

Решение – бинаризация (разбиение на бины).




Новые признаки:

- $x_{[0;10)}$ - равен 1, если

дом находится в пределах

10 км от МКАД, и 0 иначе

- $x_{[10;30)}$ - равен 1, если

дом находится в пределах от 10 км до 30 км МКАД, и 0 иначе. И т.д.

$$\boldsymbol{a(x) =}$$
$$\boldsymbol{= w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_{41} x_{[0;10)} + w_{42} x_{[10;30)} + w_{43} x_{[30;50)}}$$
$$\boldsymbol{+ w_{44} x_{\geq 50}}$$