



Онлайн-образование

Меня хорошо видно && слышно?

Ставьте + , если все хорошо
Напишите в чат, если есть проблемы

Проверить, идет ли запись!



«Логистическая регрессия»



Андрей Канашов

Data Scientist
OMD OM GROUP
@Андрей Канашов

Преподаватель



Андрей Канашов

- Data Scientist в OMD OM GROUP
 - Кластерный анализ целевых аудиторий
 - Персонализация рекламы
 - Анализ социальных сетей

Правила вебинара



Активно участвуем



Задаем вопрос в чат или голосом



Off-topic обсуждаем в Slack #канал группы или #general



Вопросы вижу в чате, могу ответить не сразу

Цели вебинара | После занятия вы узнаете

1

Логистическая регрессия

2

Градиентный спуск

3

ROC-AUC

4

Применение на практике

Задачи машинного обучения

Задачи машинного обучения

С учителем
Supervised learning



Классификация



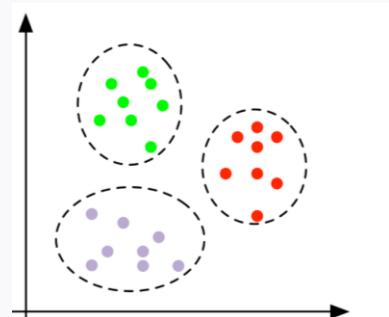
Регрессия



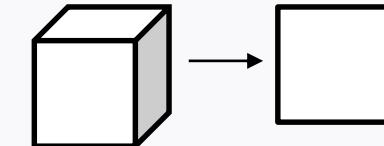
Без учителя
Unsupervised learning



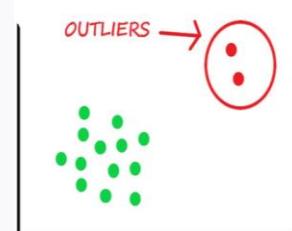
Кластеризация



Снижение
размерности



Поиск
аномалий

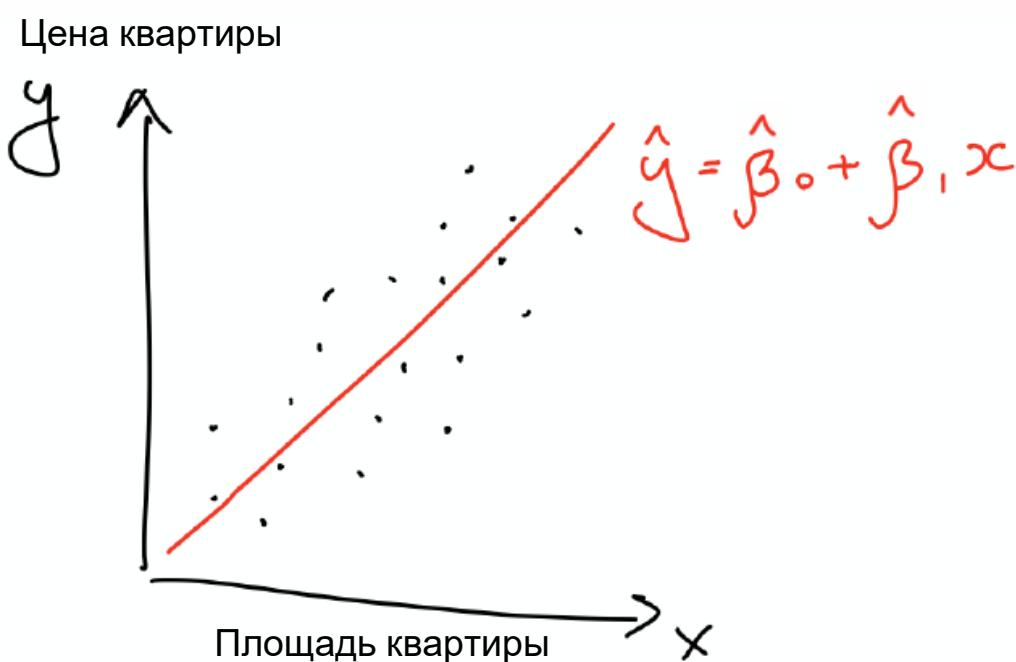


Логистическая регрессия

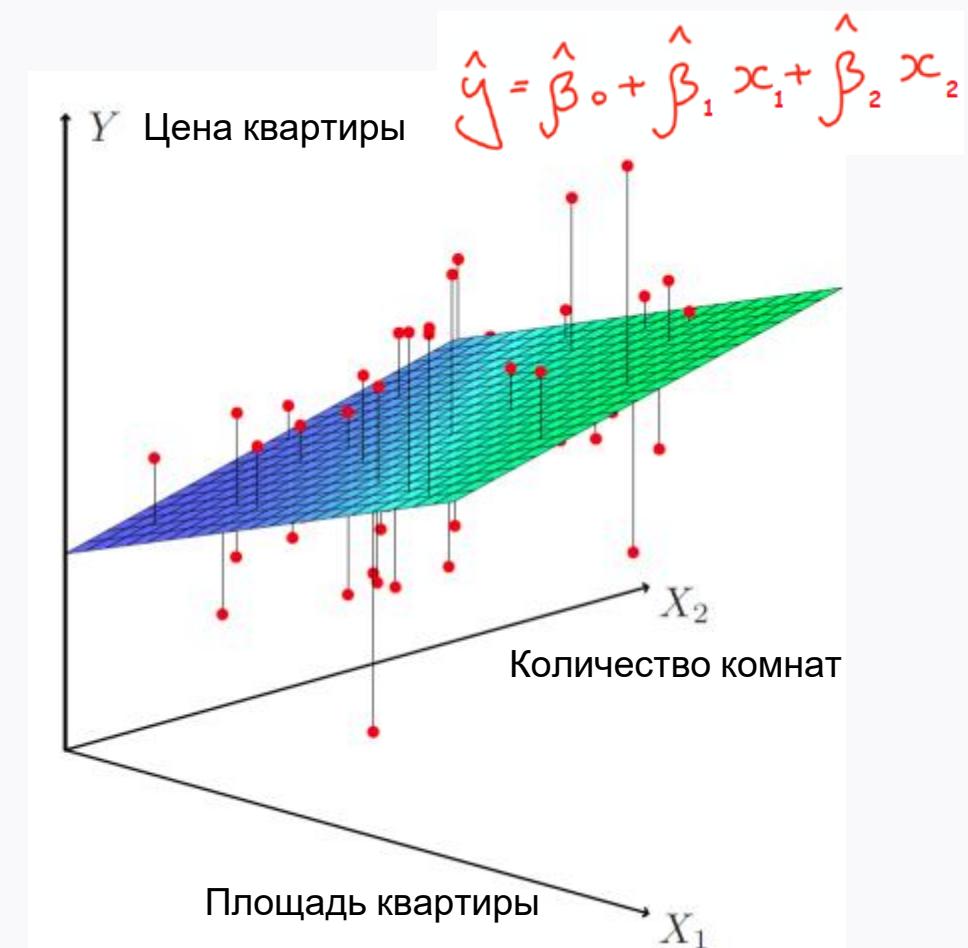
Регрессия vs Классификация

Регрессия

один признак x



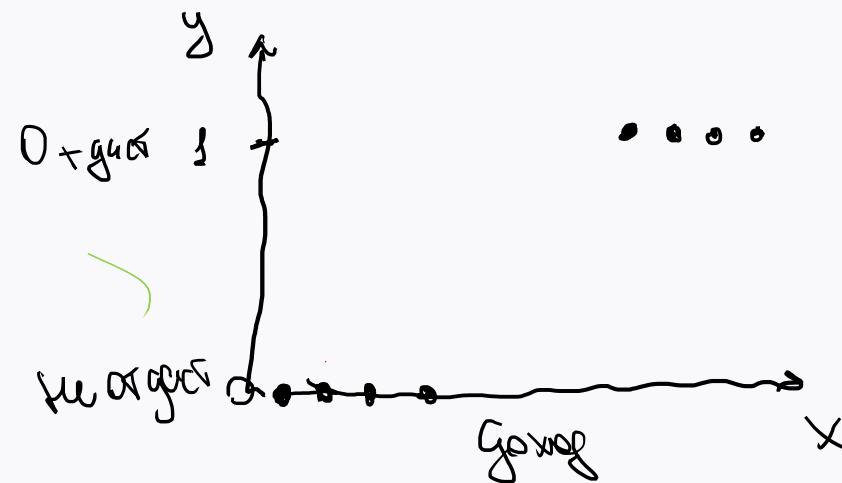
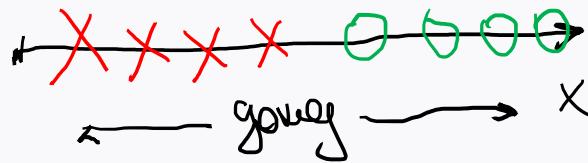
два признака x_1 и x_2



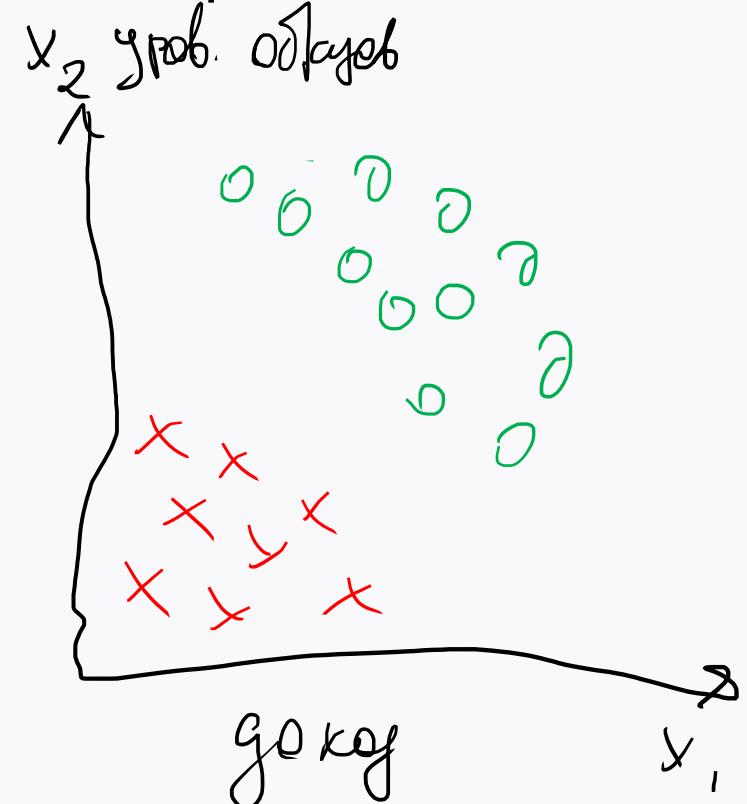
Регрессия vs Классификация

Классификация

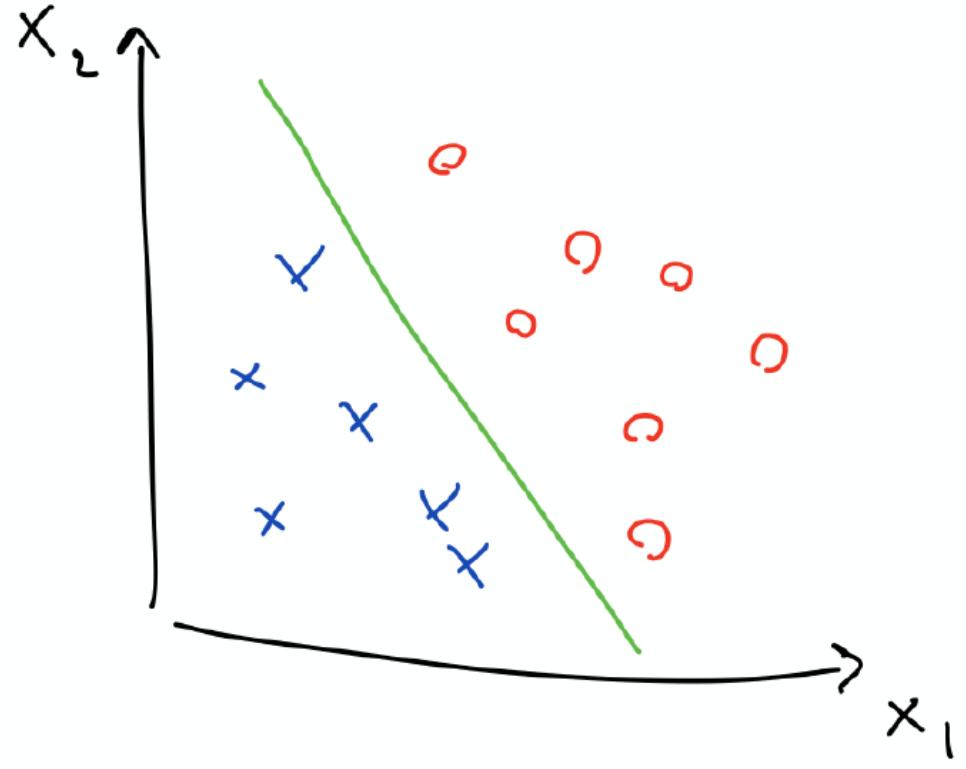
один признак x



два признака x_1 и x_2



Линейный классификатор



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

В каких пределах может изменяться \hat{y} ?

$$\hat{y} \in (-\infty; +\infty)$$

Нужны метки $\hat{y} \in [0; 1]$

Будем предсказывать:

$$\hat{y} = 1, \text{ если } \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \geq 0$$

$$\hat{y} = 0, \text{ если } \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 < 0$$

В общем случае:

$$\hat{y} = sign(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_n x_n) = sign(\beta^T x)$$

Логистическая регрессия

Логистическая регрессия предсказывает вероятность принадлежности к определенному классу.

$$0 \leq \hat{y} \leq 1$$

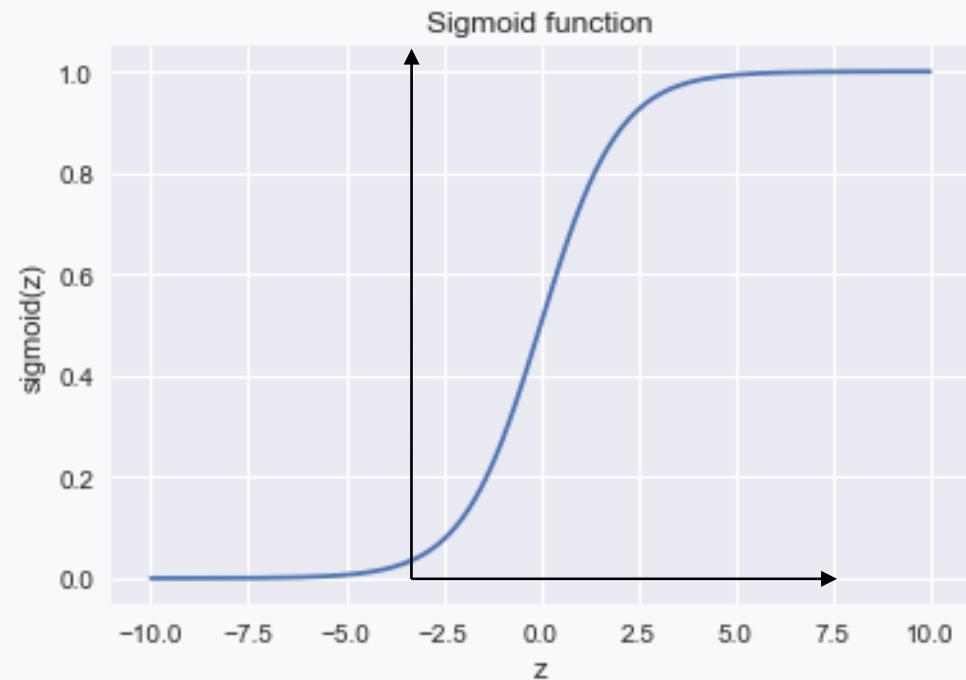
$$\hat{y} = f(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_n x_n) = f(\beta^T x)$$

Сигмоидная функция (логистическая функция):

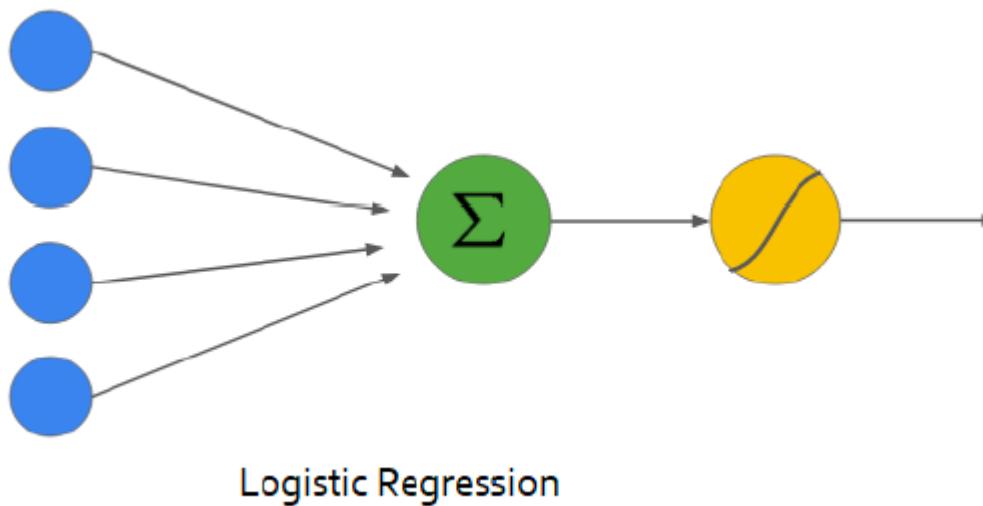
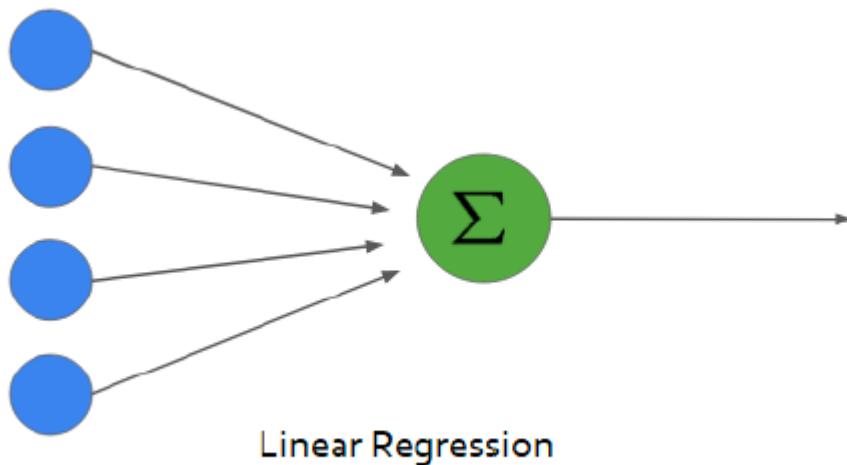
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\sigma(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_n x_n) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_n x_n)}}$$

$$\sigma(\beta^T x) = \frac{1}{1 + e^{-\beta^T x}} = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$



Логистическая регрессия



Логистическая регрессия

$P(X)$ - вероятность происходящего события X
[0; 1]

$$P(Y = 1 | x_1, \dots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}$$

Отношение шансов (odds ratio):

$$OR(X) = \frac{P(X)}{1 - P(X)} \quad \text{- отношение вероятностей того, произойдет ли событие или не произойдет}$$

[0; +\infty)



Логарифм отношения шансов:

$$\text{logit}(X) = \log(OR(X)) = \log\left(\frac{P(X)}{1 - P(X)}\right)$$

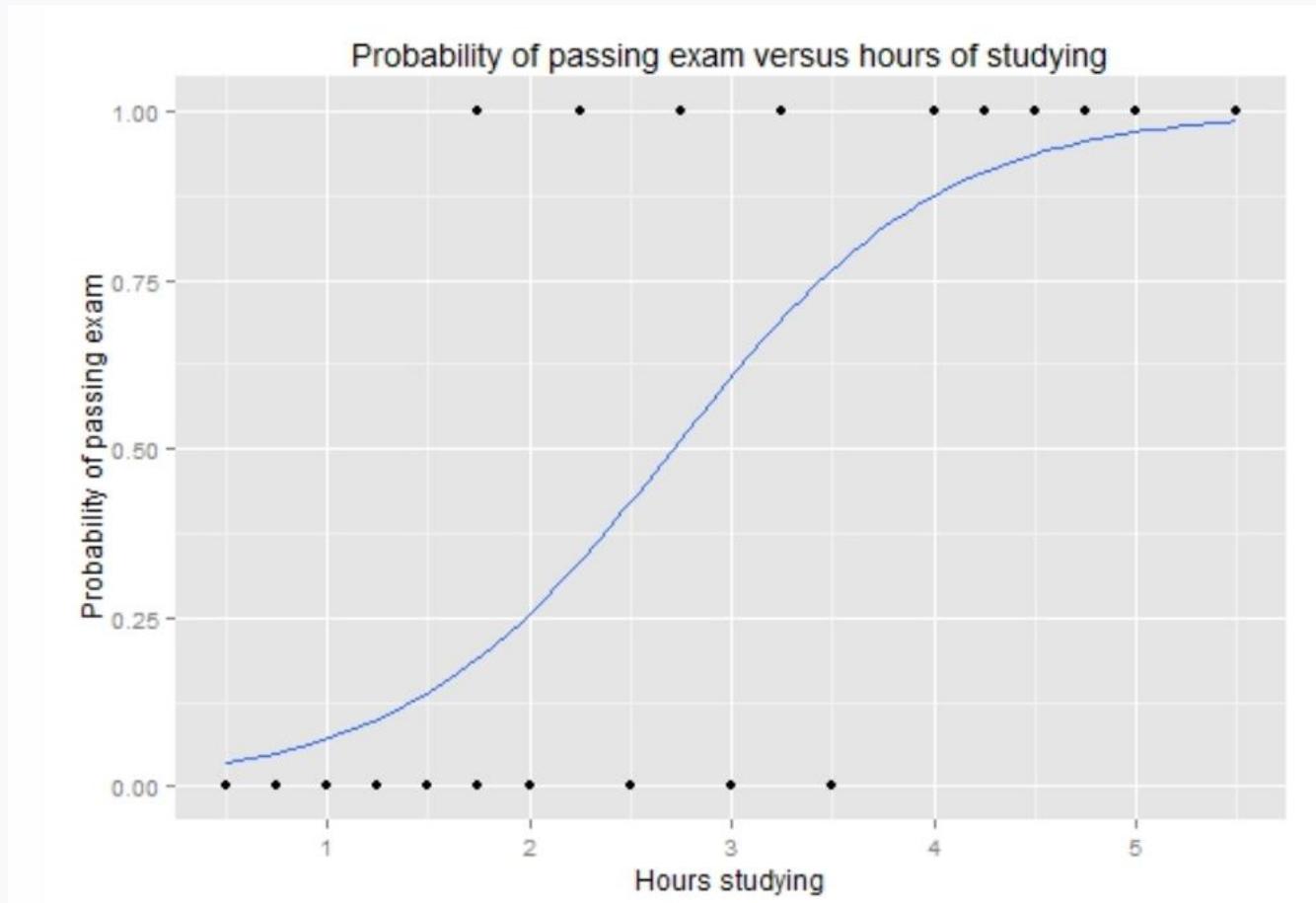
(-\infty; +\infty)

$$\text{logit}(P(Y = 1 | x_1, \dots, x_k)) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

$$\log\left(\frac{P(X)}{1 - P(X)}\right) = \beta^T x \rightarrow P(X) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$

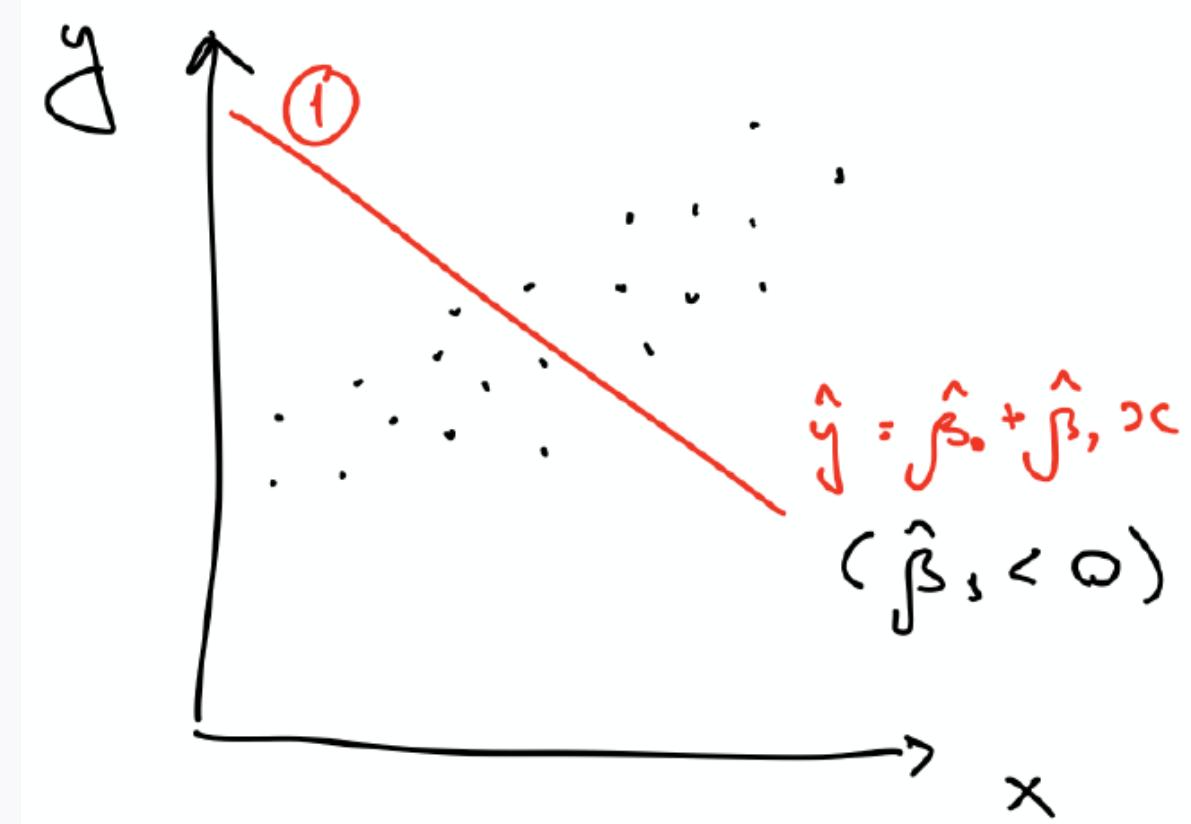
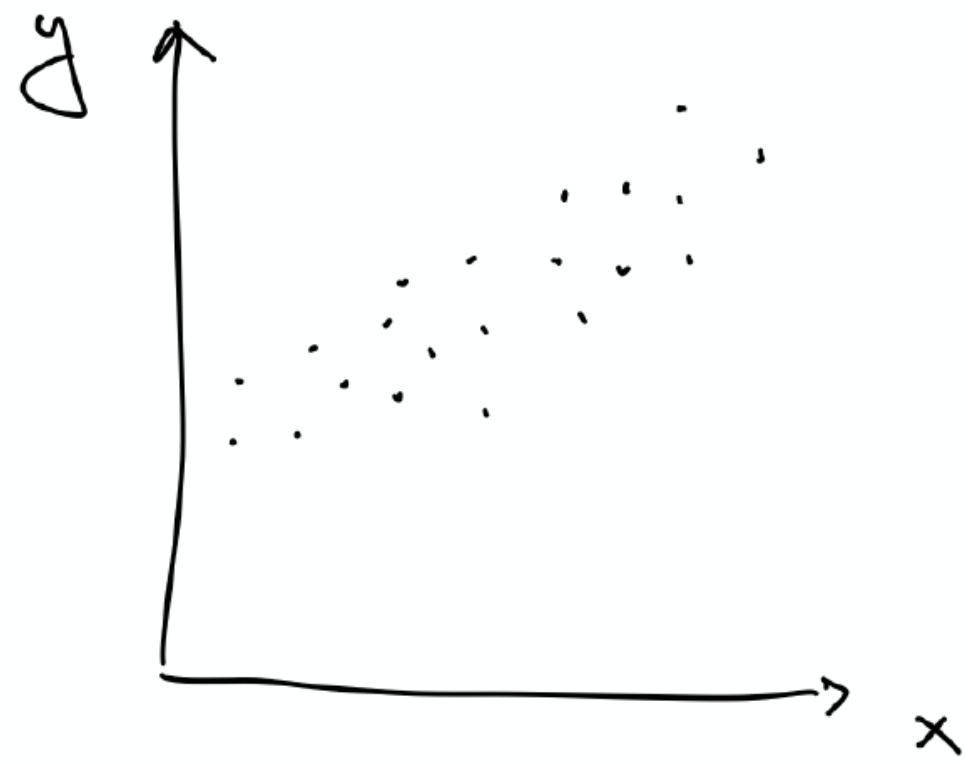
В модели логистической регрессии граничная функция определяет логарифм отношения шансов класса

Логистическая регрессия - пример

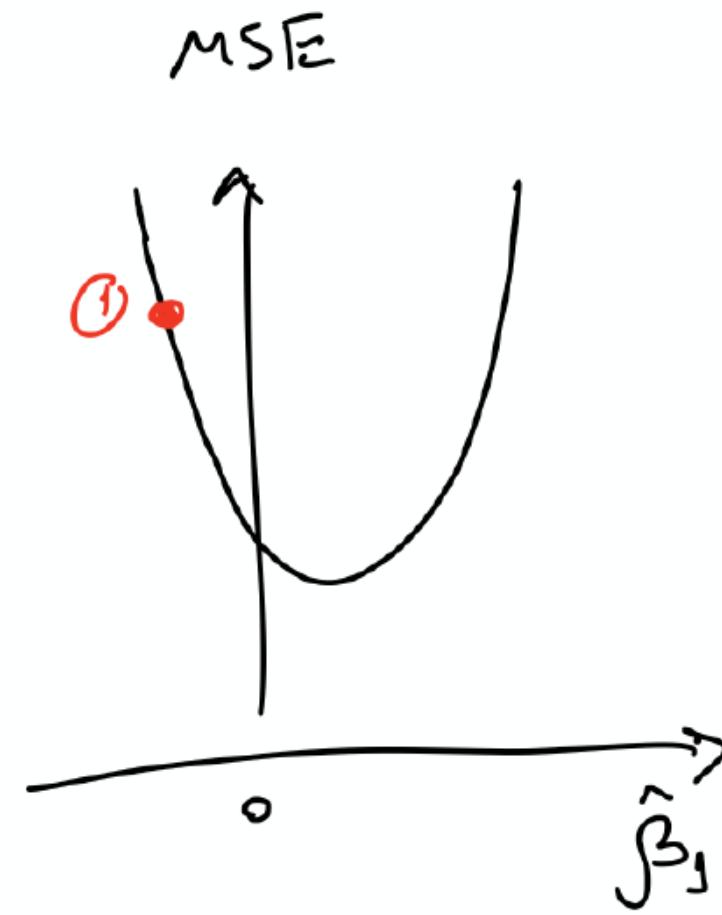
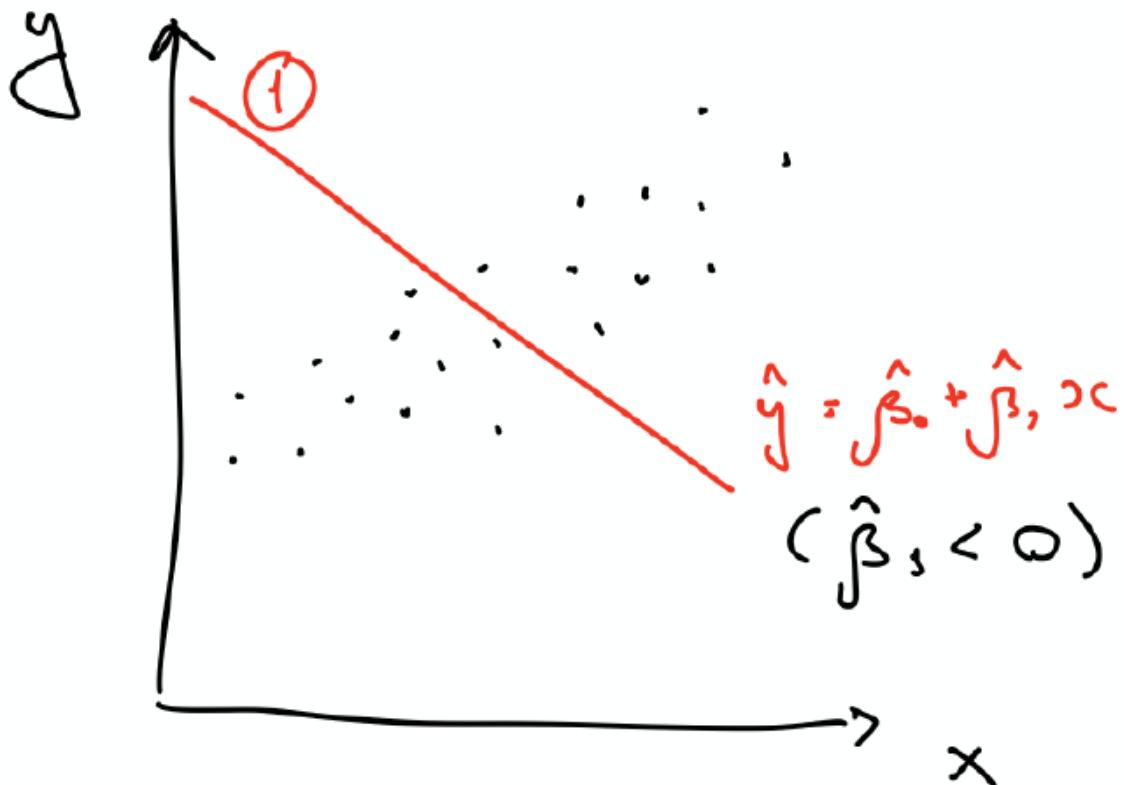


Градиентный спуск

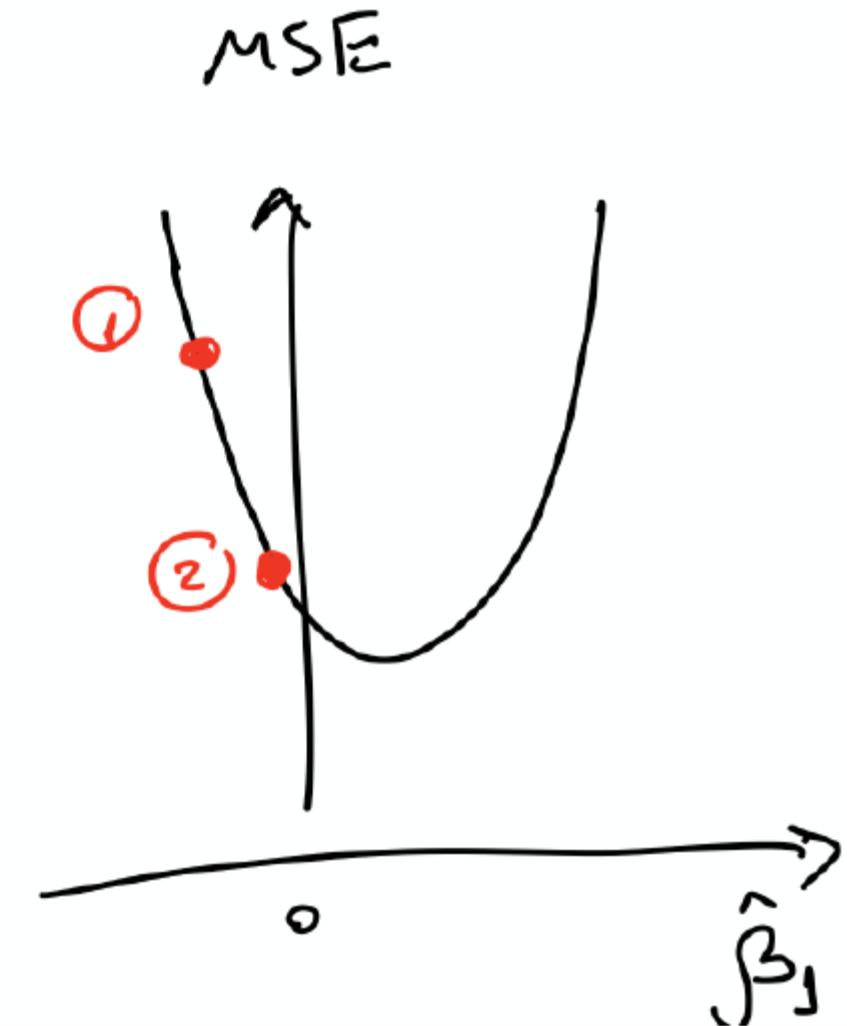
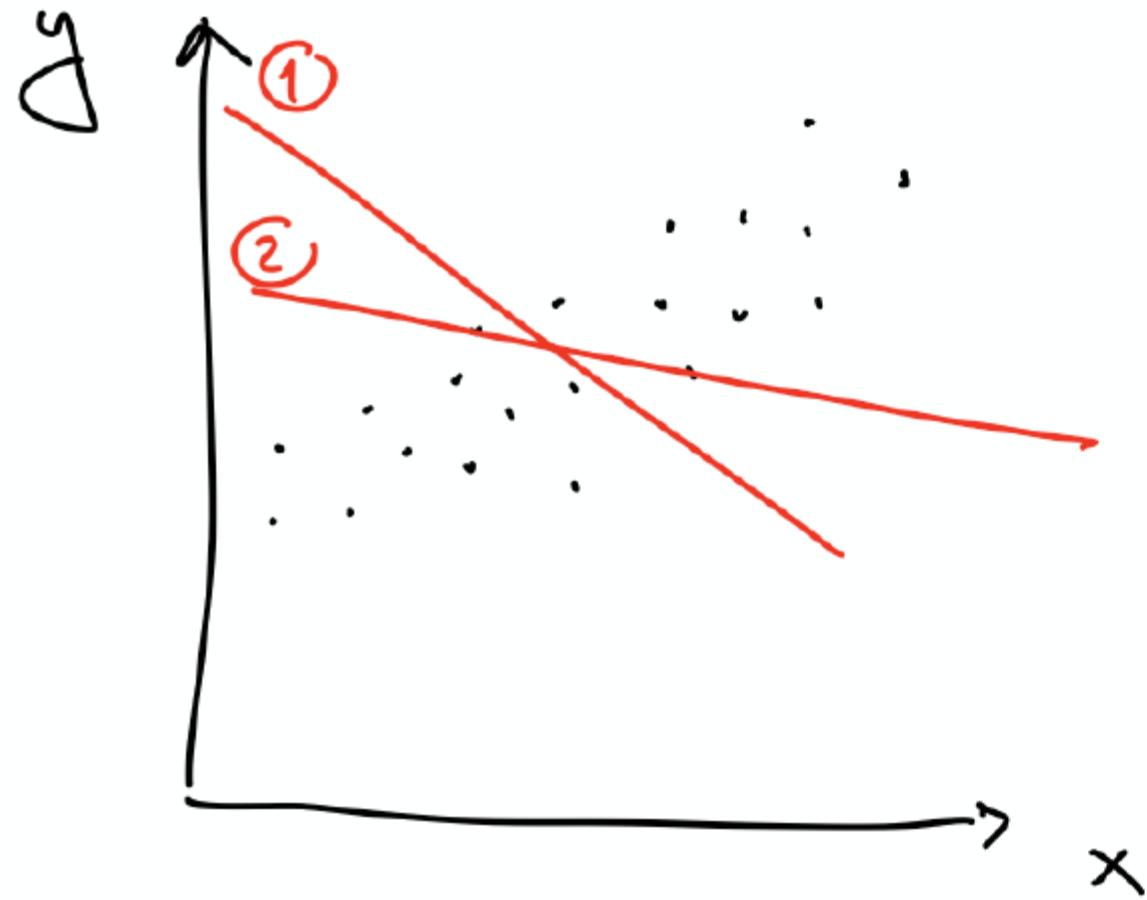
Градиентный спуск



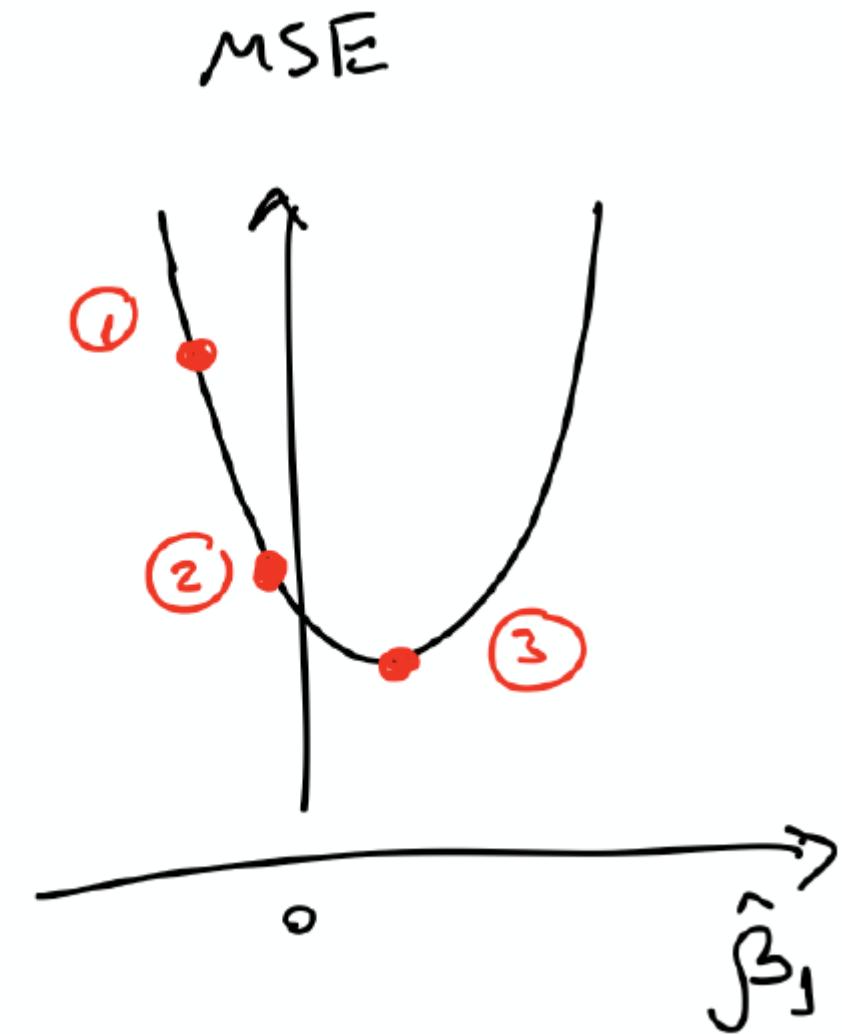
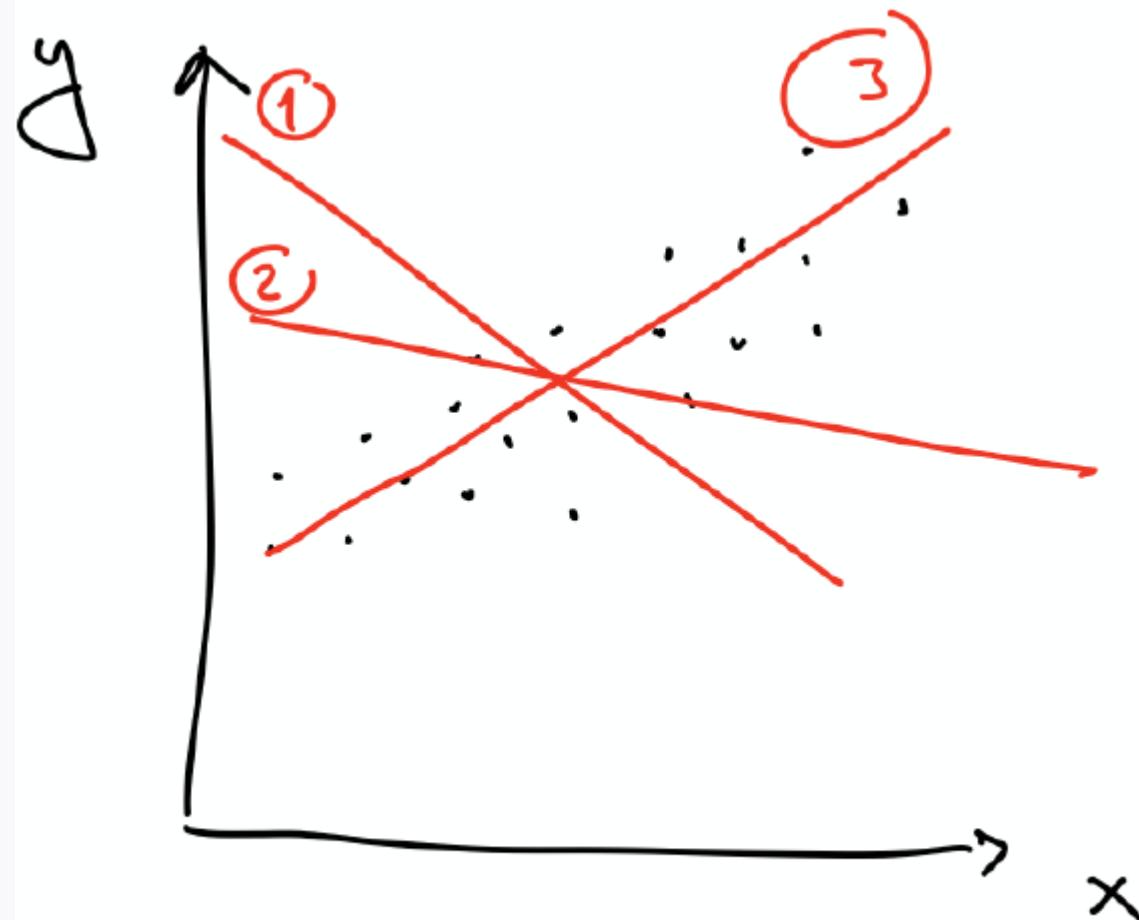
Градиентный спуск



Градиентный спуск



Градиентный спуск



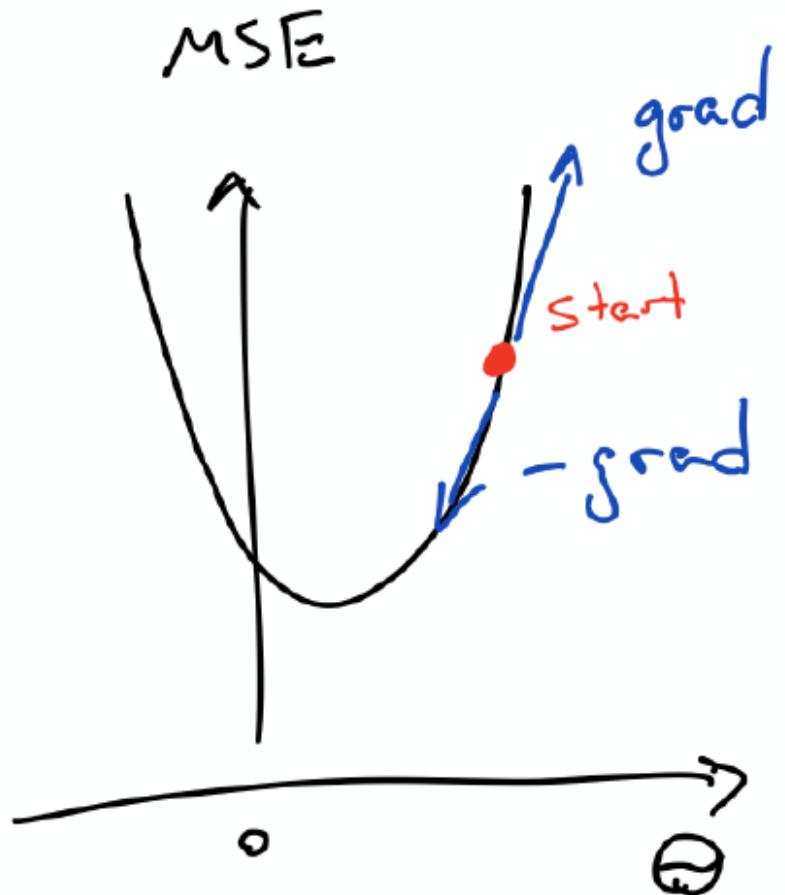
Градиентный спуск

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\frac{\partial MSE}{\partial \theta} = \frac{1}{n} \sum_i (y_i - \theta^T x_i)^2$$

$$= \frac{2}{n} \sum_i (y_i - \theta^T x_i) \cdot x_i$$

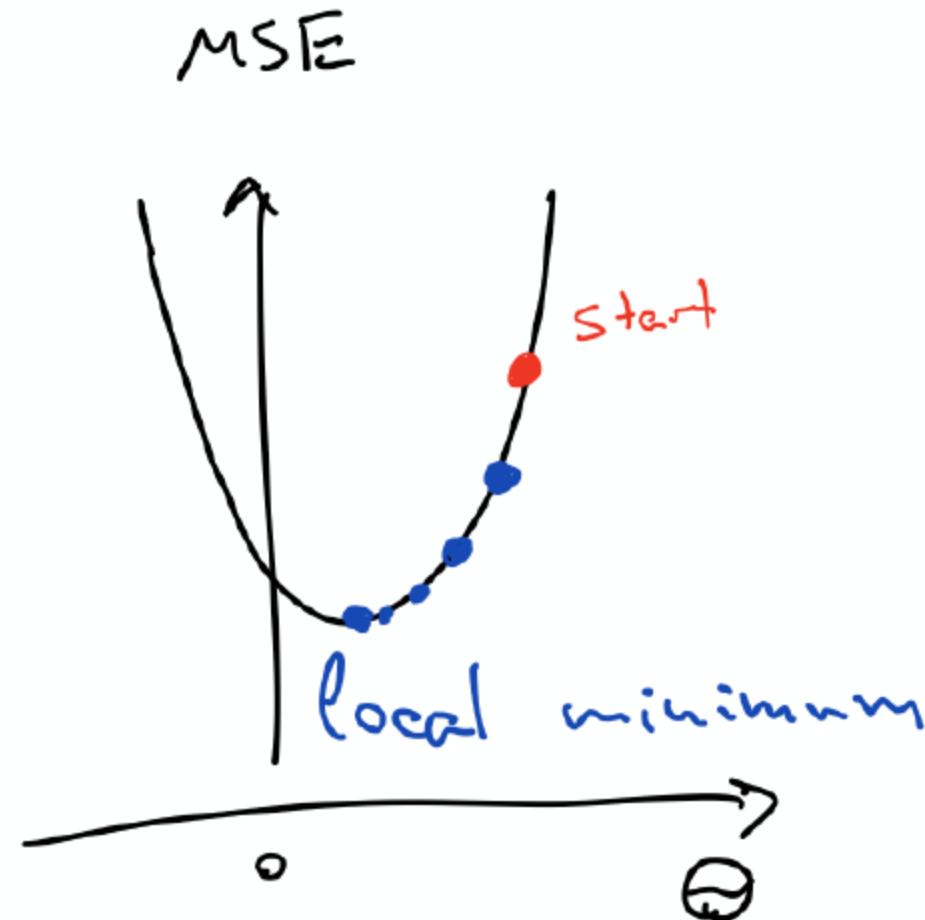
Градиентный спуск



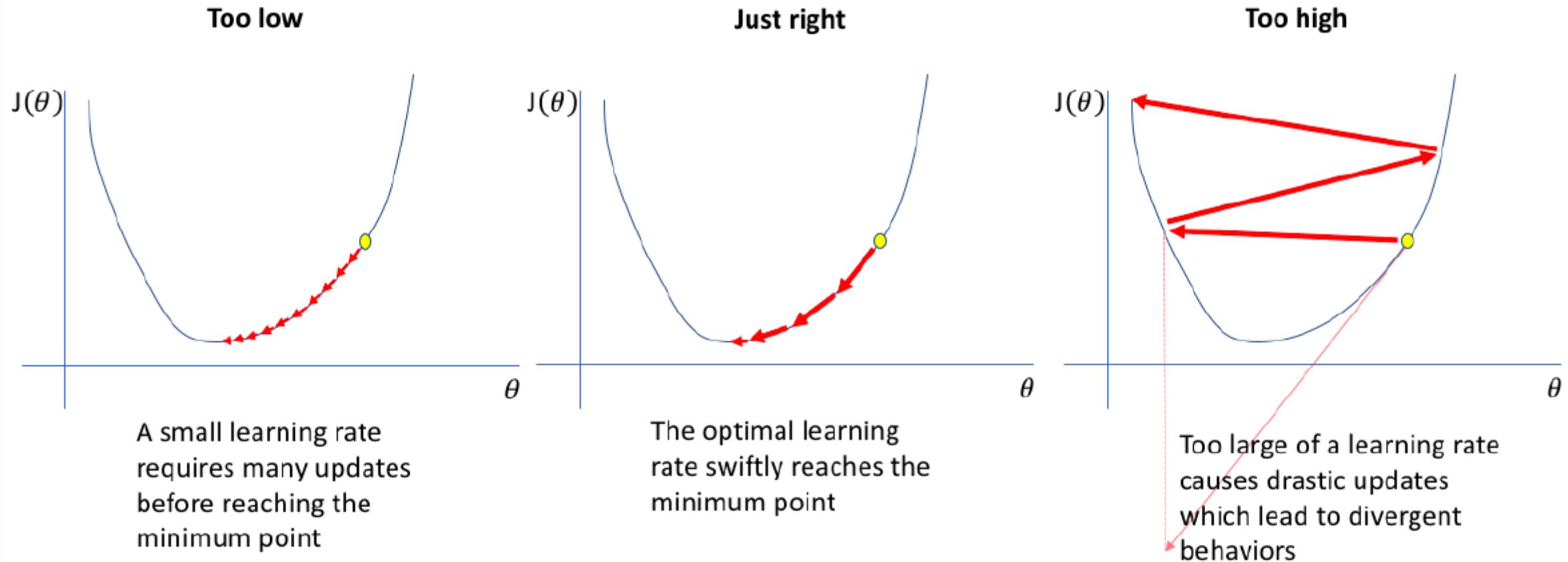
$$\theta_i := \theta_i - \alpha \text{grad}$$

$$\theta_i := \theta_i - \alpha \frac{1}{n} \sum (y_i - \hat{y}) x_i$$

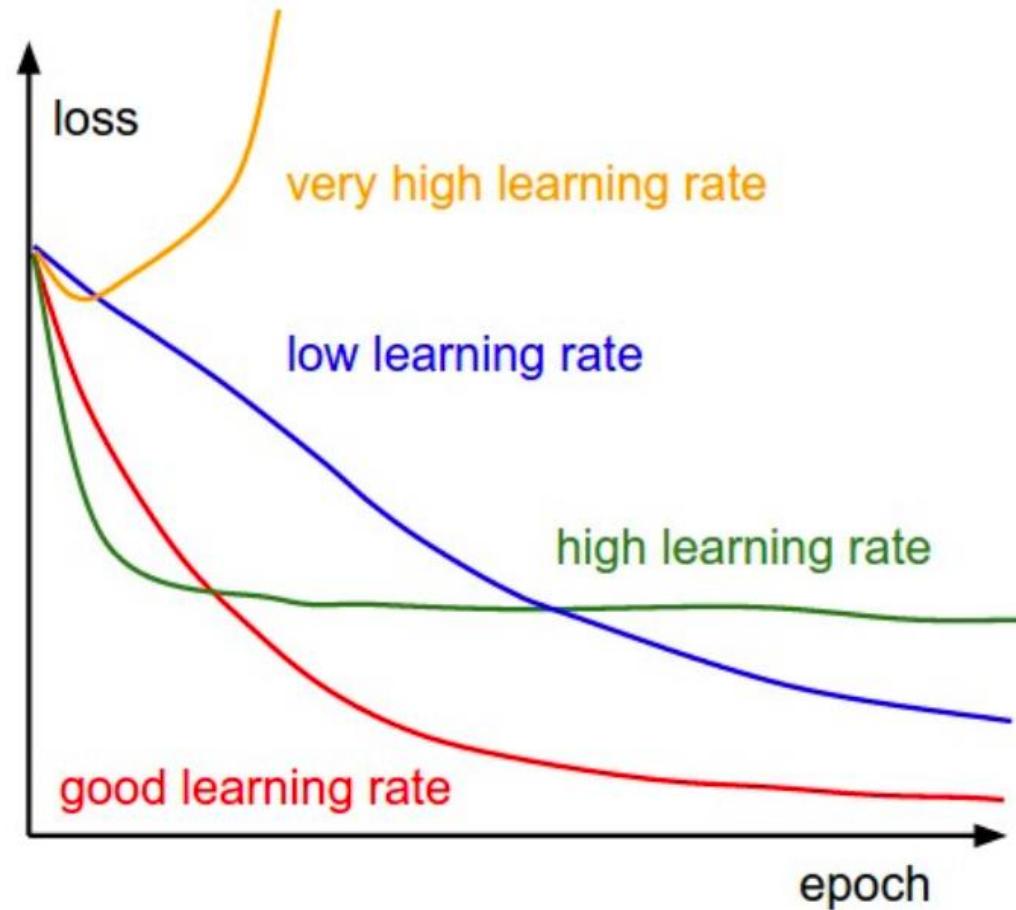
Градиентный спуск



Learning rate

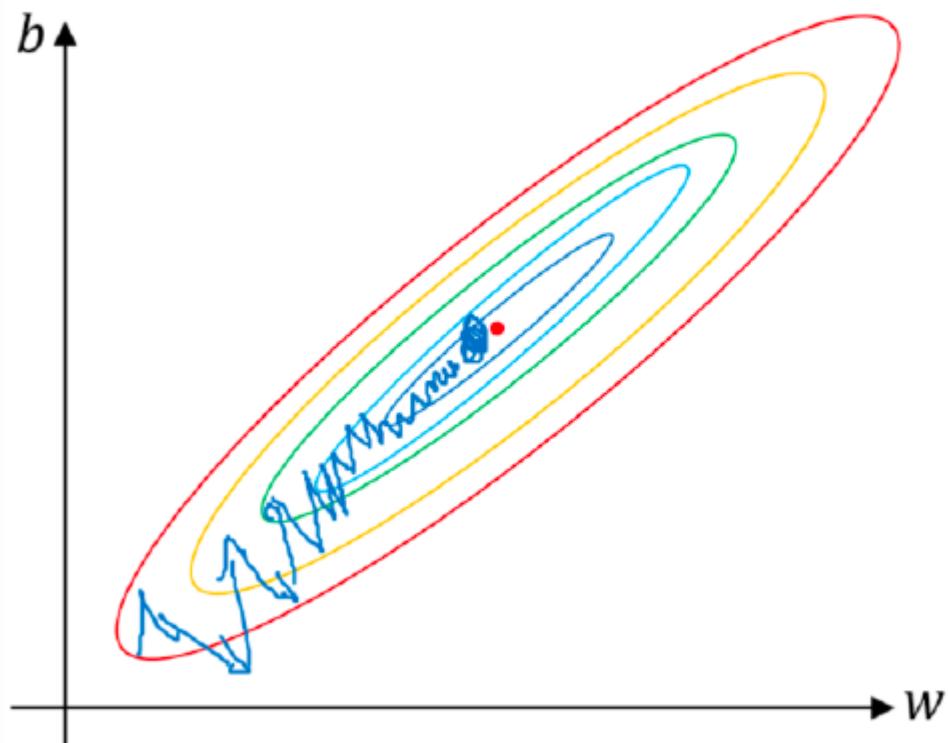


Learning rate

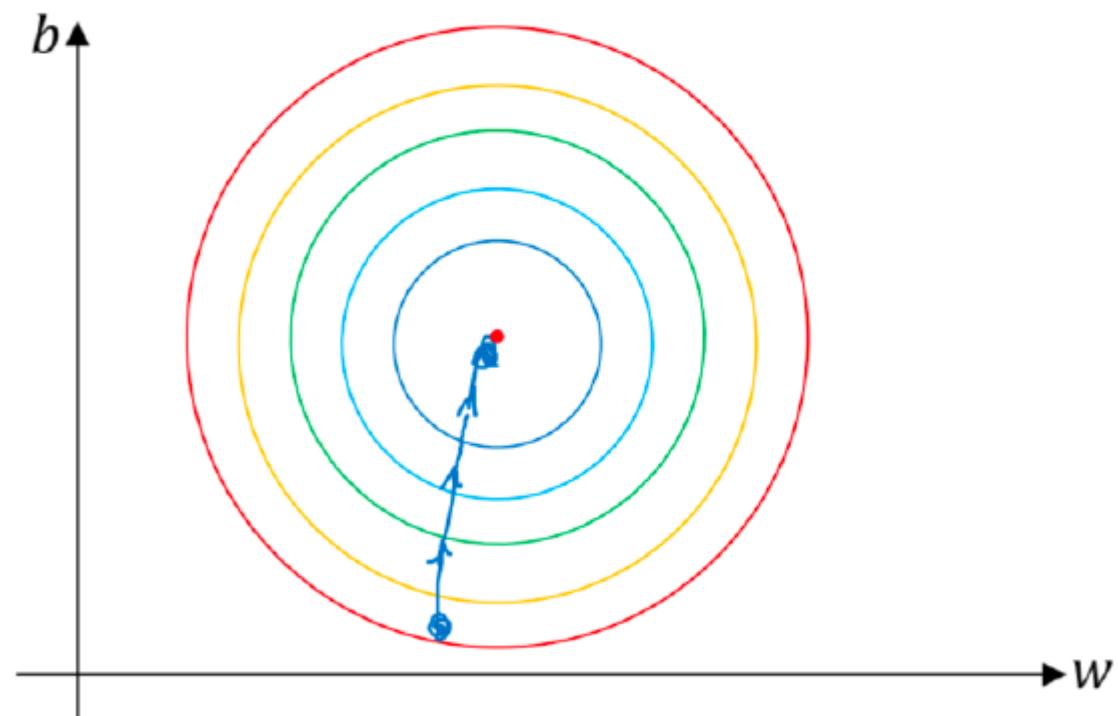


Нормализация данных

Unnormalized



Normalized



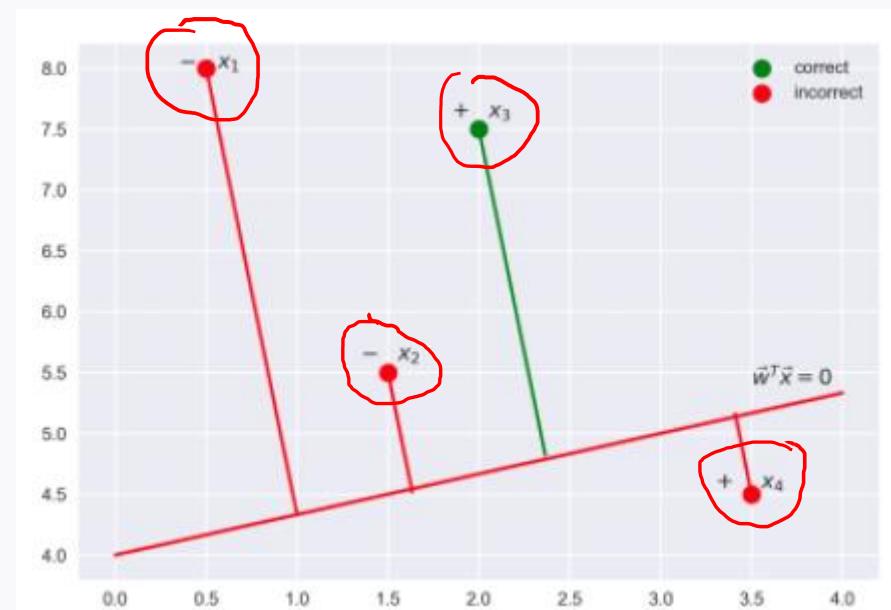


LIVE

Отступ

Значит, выражение $M(\vec{x}_i) = y_i \vec{w}^T \vec{x}_i$ – это своего рода "уверенность" модели в классификации объекта x_i :

- если отступ большой (по модулю) и положительный, это значит, что метка класса поставлена правильно, а объект находится далеко от разделяющей гиперплоскости (такой объект классифицируется уверенно). На рисунке – x_3 .
- если отступ большой (по модулю) и отрицательный, значит метка класса поставлена неправильно, а объект находится далеко от разделяющей гиперплоскости (скорее всего такой объект – аномалия, например, его метка в обучающей выборке поставлена неправильно). На рисунке – x_1 .
- если отступ малый (по модулю), то объект находится близко к разделяющей гиперплоскости, а знак отступа определяет, правильно ли объект классифицирован. На рисунке – x_2 и x_4 .





LIVE

Плюсы и минусы логистической регрессии

Плюсы:

- + Очень простая и быстрая в обучении, предсказании
- + Может не только в линейные зависимости
- + Легко интерпретировать

Первое, что стоит пробовать в задаче классификации :)

Минусы:

- Не всегда предпосылка о линейной зависимости выполняется в данных
- Может быть слишком простой моделью для ваших данных
- Крайне чувствительна к предобработке данных

Проверка достижения целей

Цели вебинара | Проверка достижения целей

1

Что предсказывает логистическая регрессия?

2

Какие параметры в логистической регрессии?

3

Как работает градиентный спуск?

4

Какие метрики?

Рефлексия



Достигли ли вы цели вебинара?



С какими основными мыслями и инсайтами
уходите с вебинара?

Следующий вебинар

Тема: «Feature engineering & advanced preprocessing»



Пятница 25 сентября 20:00



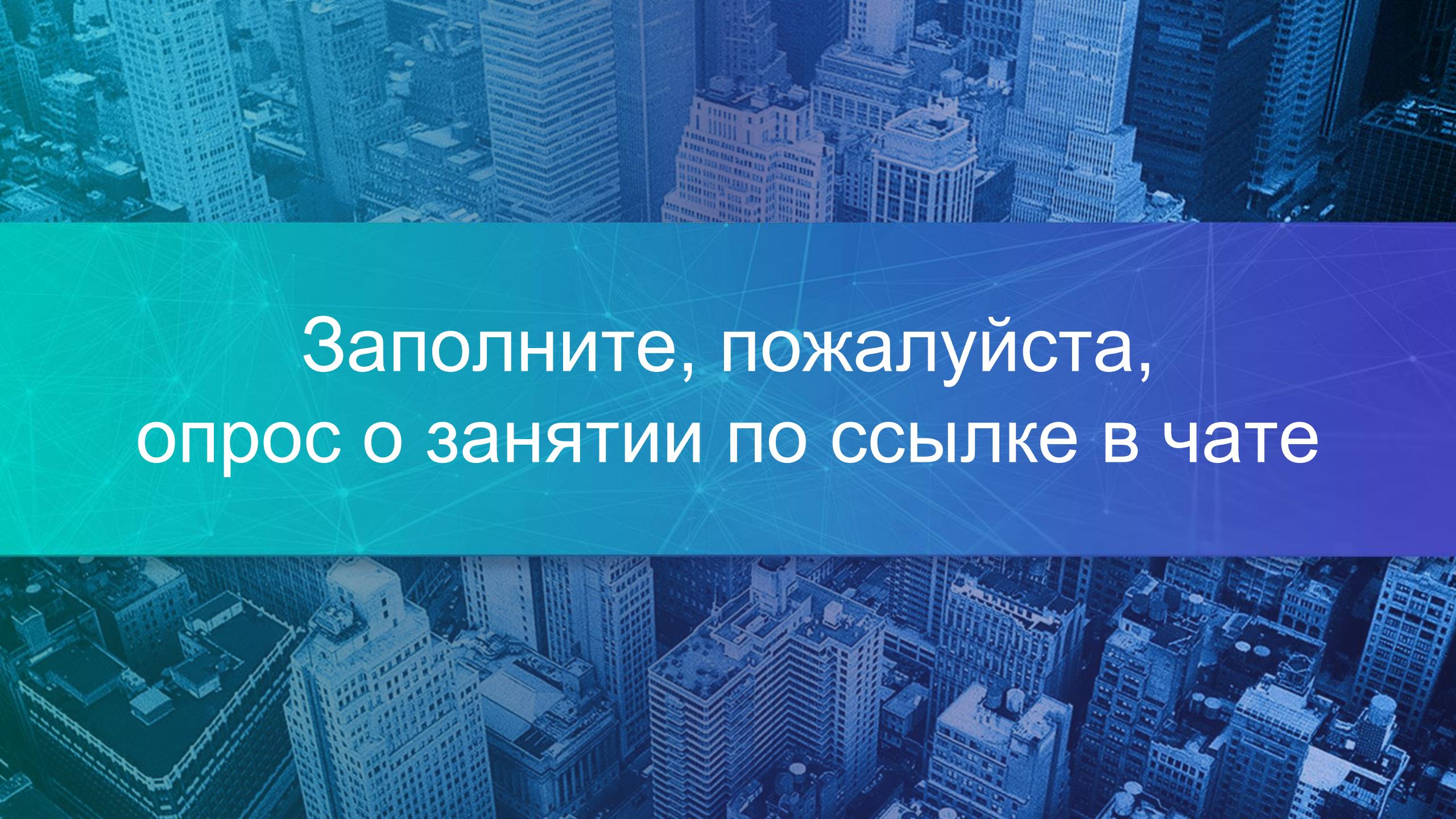
Ссылка на вебинар будет в ЛК за 15 минут



Материалы к занятию
в ЛК — можно изучать



Обязательный
материал обозначен
красной лентой



Заполните, пожалуйста,
опрос о занятии по ссылке в чате

Спасибо за внимание!
Приходите на следующие вебинары



Андрей Канашов
Data Scientist
OMD OM GROUP
@Андрей Канашов