



# Задачи классификации и регрессии

Лекция 2

Иван Горбань

---

# Преподаватель



Иван Горбань

Руководитель команды

Гео и Ритейл

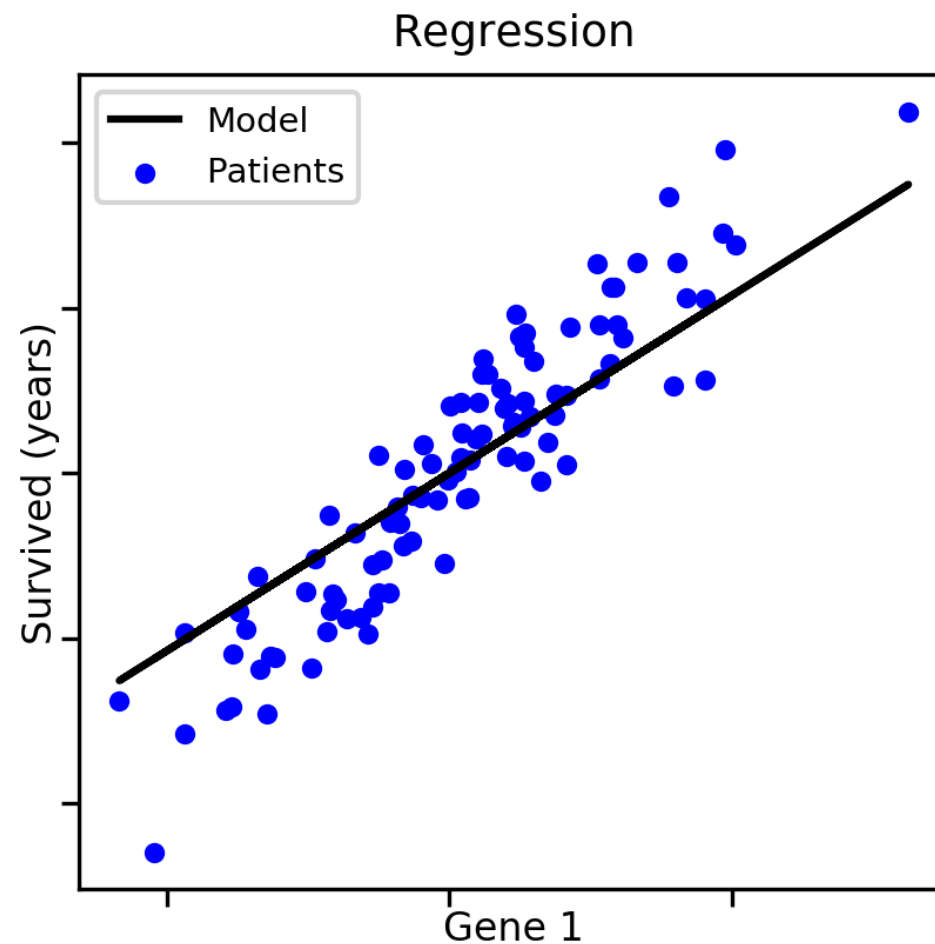
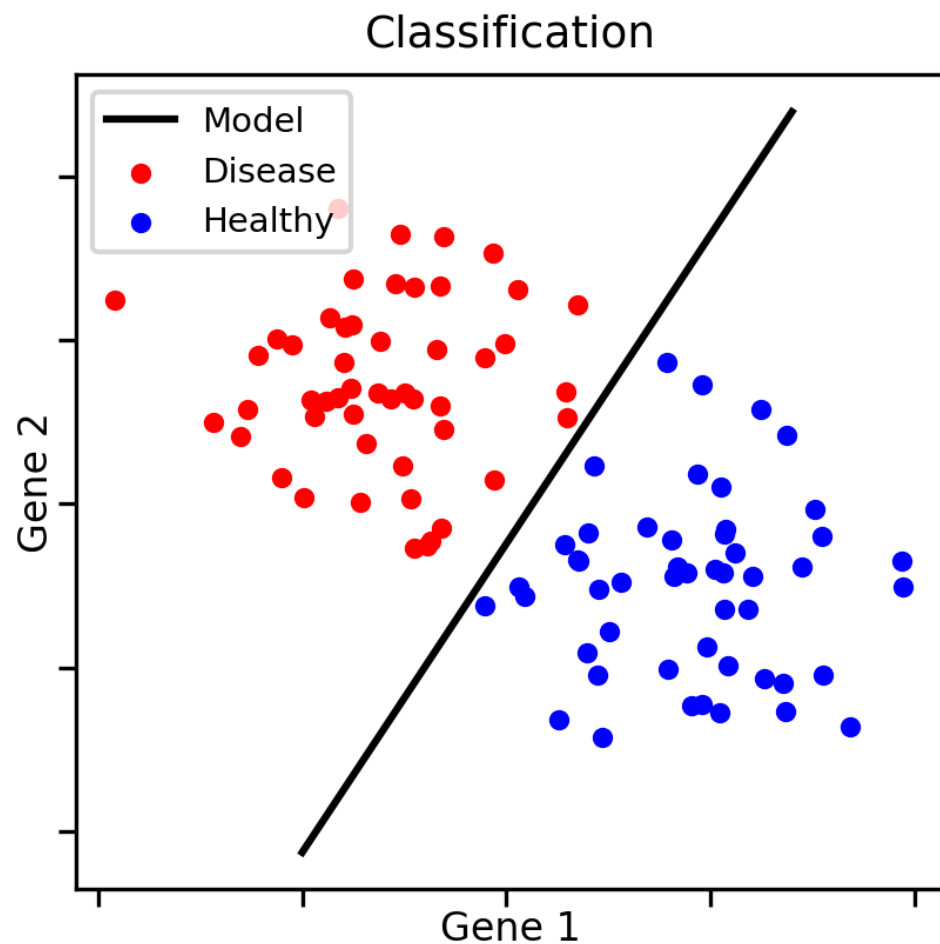
BigData MegaFon



## План занятия:

1. Задачи регрессии и классификации.
2. Bias-Variance tradeoff.
3. Методы:
  - a. kNN
  - b. Наивный Байес
  - c. Линейная регрессия
  - d. Логистическая регрессия
  - e. Регуляризация
4. Практика

# Регрессия и классификация:



# Регрессия и классификация:

- $Y = f(X) + \epsilon$  – Общая форма для записи данных. Целевая переменная является некоторой функцией от  $X$  плюс ошибка.
- $\hat{Y} = \hat{f}(X)$  – оценка целевой переменной посредством оценки функции  $f$ .
- В задаче регрессии  $Y \in \mathbb{R}$
- В задаче классификации  $Y \in \{1, \dots, M\}$

# Задача регрессии:

Признаки (Features)								Target
Площадь, м <sup>2</sup>	Число комнат	Расстояние до центра, км	Новостройка	Наличие балкона	Время до метро, мин	Этаж	Высота потолков, м	Стоимость
36	1	36	1	0	14	5	2	3 321 000
56	2	4	0	0	3	3	4	13 000 000
41	1	28	0	1	13	23	2.3	8 020 000
148	4	13	1	1	7	3	5	21 412 000
...	...	...	...	...	...	...	...	...
52	3	41	1	1	53	35	2.8	?



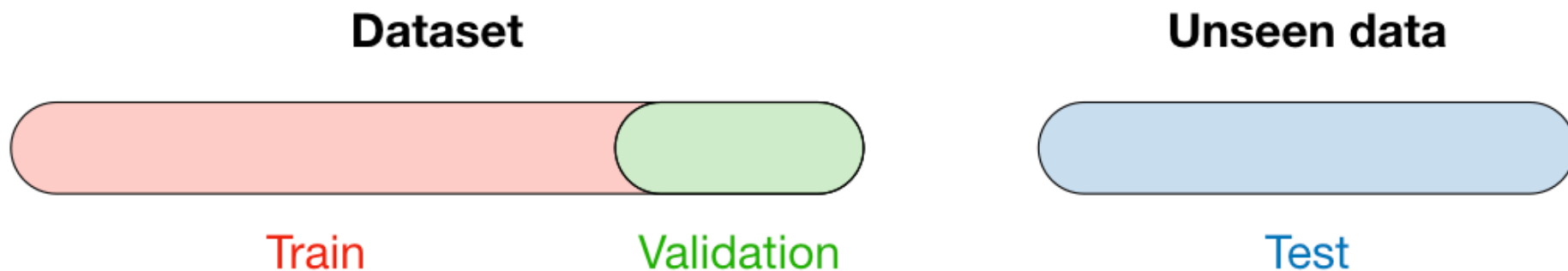
## Два слова о train-test split

В процессе построения модели данные часто делят на две или три части:

## Два слова о train-test split

В процессе построения модели данные часто делят на две или три части:

- Train
- Validation
- [Test]

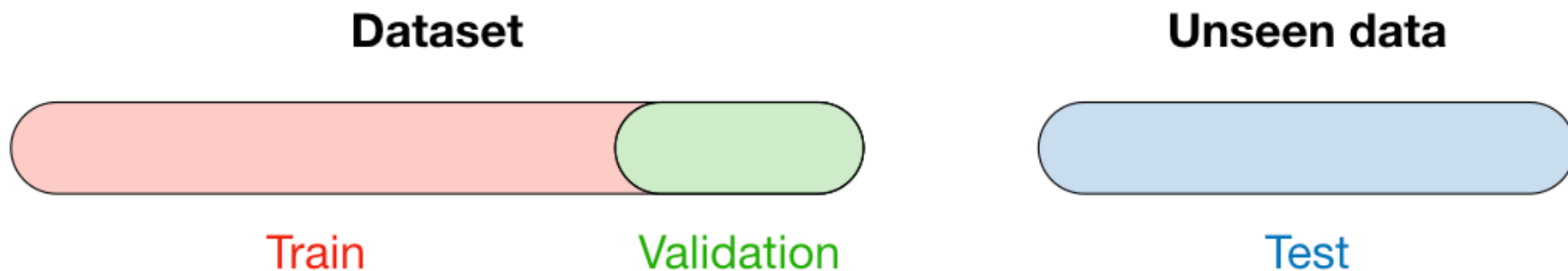




# Два слова о train-test split

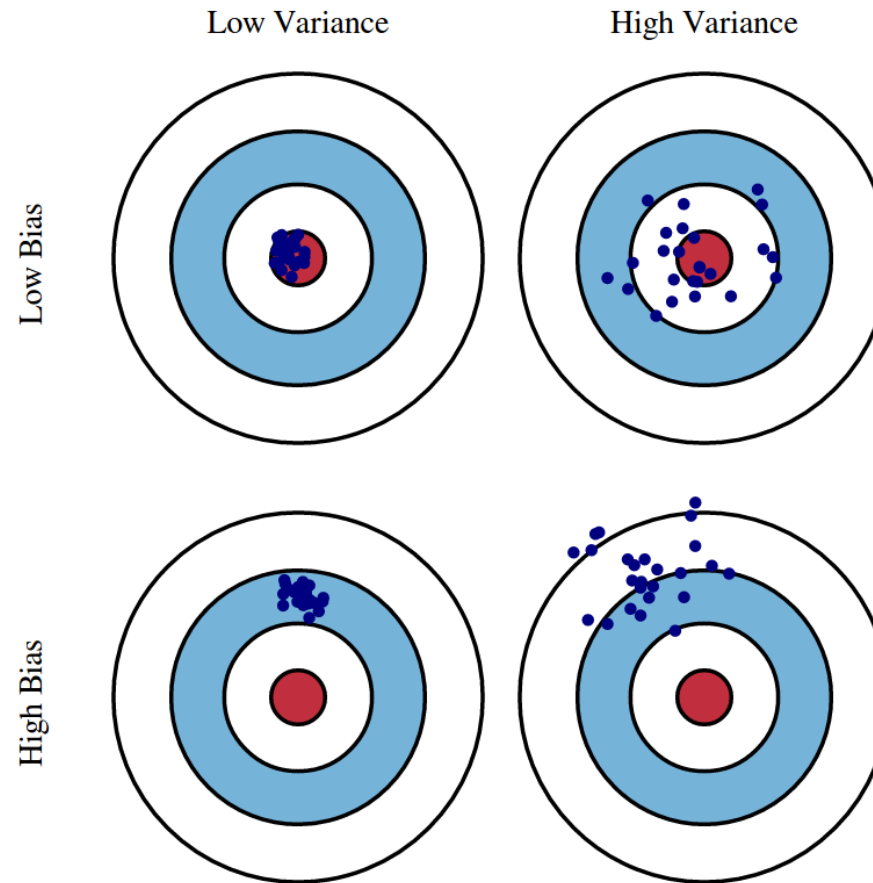
В процессе построения модели данные часто делят на две или три части:

- Train
- Validation
- [Test]

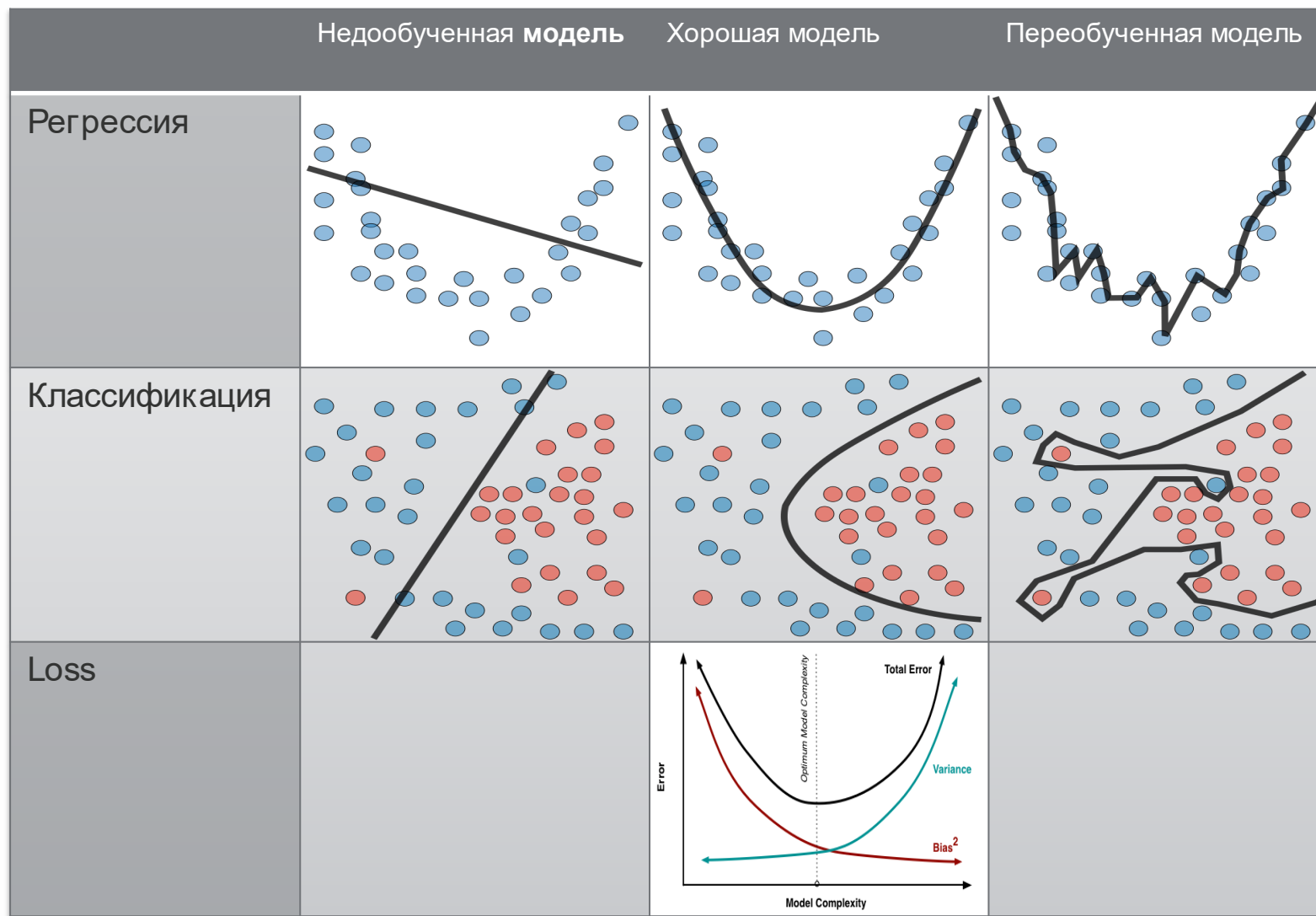


Зачем?

# Bias-variance tradeoff



# Bias-variance tradeoff



# Nearest Neighbors

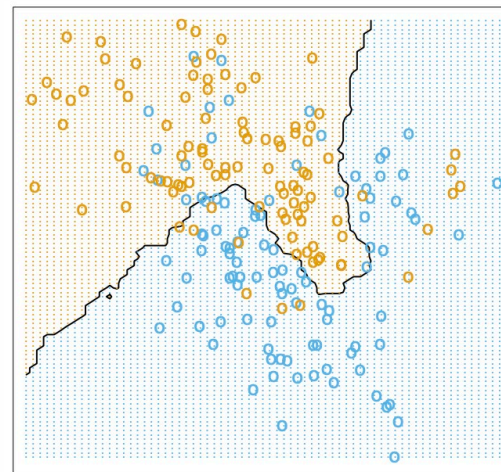
Непараметрический метод, требующий минимальное число предположений о характере функции  $f$ .

- Для регрессии:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

- Для классификации:

$$\hat{p}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$



---

# Nearest Neighbors

- Какие проблемы могут возникнуть?

---

# Nearest Neighbors

- Какие проблемы могут возникнуть?
- Выбор  $k$

---

# Nearest Neighbors

- Какие проблемы могут возникнуть?
- Выбор  $k$
- Выбор метрики близости

---

# Nearest Neighbors

- Какие проблемы могут возникнуть?
- Выбор  $k$
- Выбор метрики близости
- Скорость вычислений

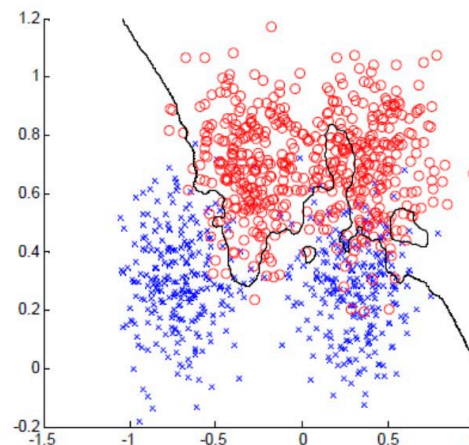
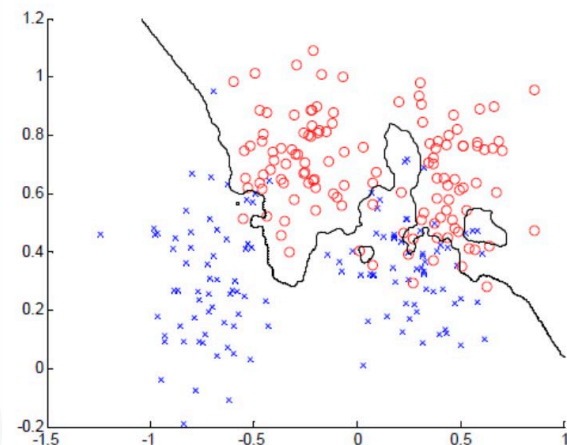
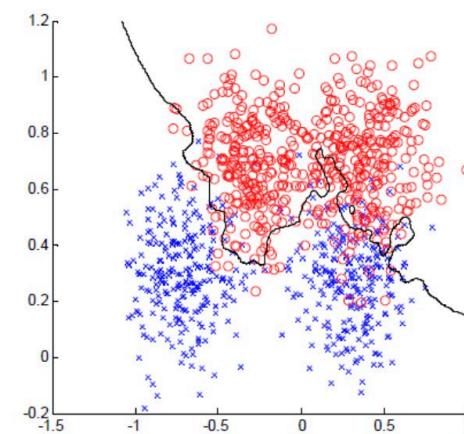
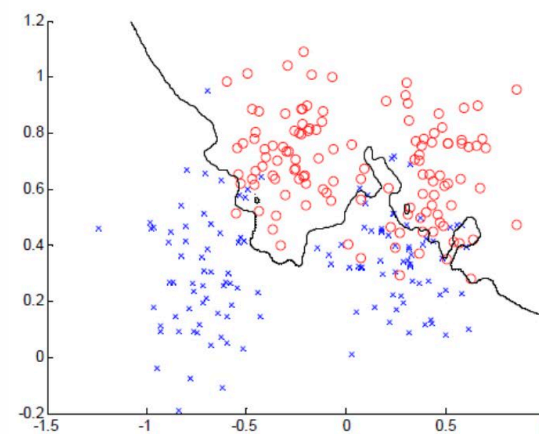
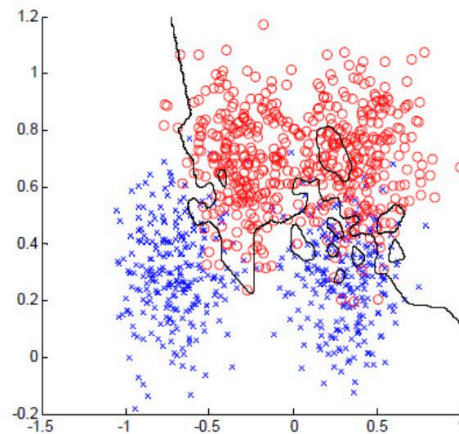
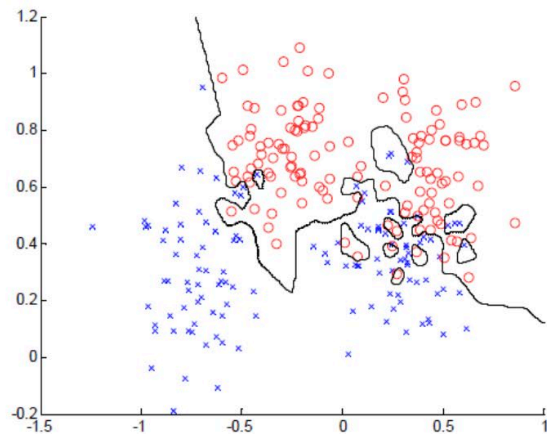


---

# Nearest Neighbors

- Какие проблемы могут возникнуть?
- Выбор  $k$
- Выбор метрики близости
- Скорость вычислений
- Проклятие размерности

# Nearest Neighbors



Какие значения  $k$  в данных примерах классификации?

# Nearest Neighbors

Наиболее частым является использование **Minkowski distance** в качестве меры близости объектов:

$$\text{dist}(\mathbf{x}, \mathbf{z}) = \left( \sum_{r=1}^d |x_r - z_r|^p \right)^{1/p}$$

# Nearest Neighbors

Наиболее частым является использование **Minkowski distance** в качестве меры близости объектов:

$$\text{dist}(\mathbf{x}, \mathbf{z}) = \left( \sum_{r=1}^d |x_r - z_r|^p \right)^{1/p}$$

Как называется данная метрика:

- При  $p=1$ ?

# Nearest Neighbors

Наиболее частым является использование **Minkowski distance** в качестве меры близости объектов:

$$\text{dist}(\mathbf{x}, \mathbf{z}) = \left( \sum_{r=1}^d |x_r - z_r|^p \right)^{1/p}$$

Как называется данная метрика:

- При  $p=1$ ?
- При  $p=2$ ?

# Nearest Neighbors

Наиболее частым является использование **Minkowski distance** в качестве меры близости объектов:

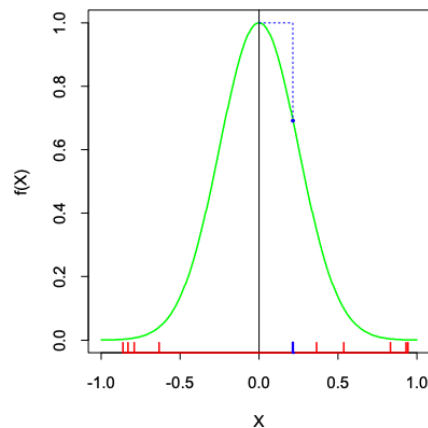
$$\text{dist}(\mathbf{x}, \mathbf{z}) = \left( \sum_{r=1}^d |x_r - z_r|^p \right)^{1/p}$$

Как называется данная метрика:

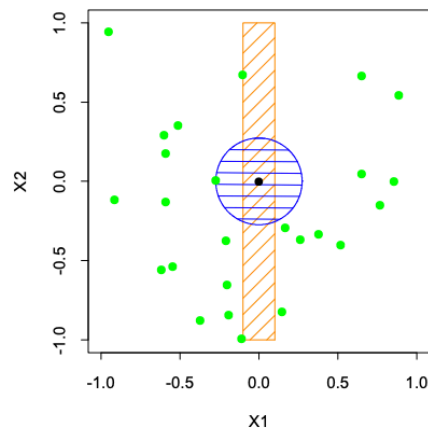
- При  $p=1$ ?
- При  $p=2$ ?
- При  $p \rightarrow \infty$ ?

# Nearest Neighbors

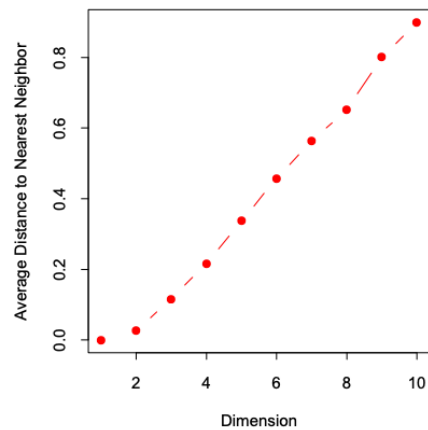
1-NN in One Dimension



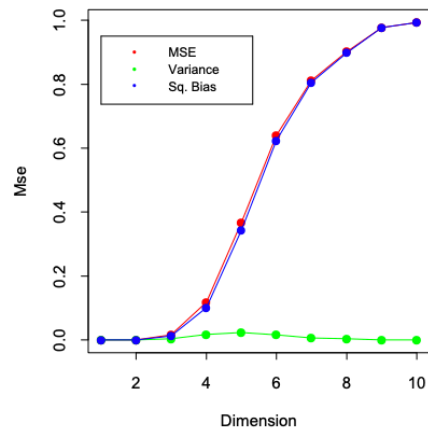
1-NN in One vs. Two Dimensions



Distance to 1-NN vs. Dimension



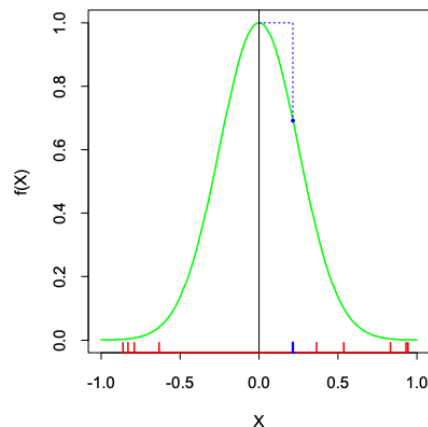
MSE vs. Dimension



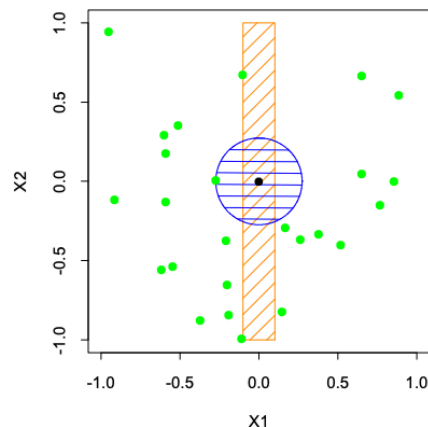
В чем состоит **проклятие размерности** для kNN?

# Nearest Neighbors

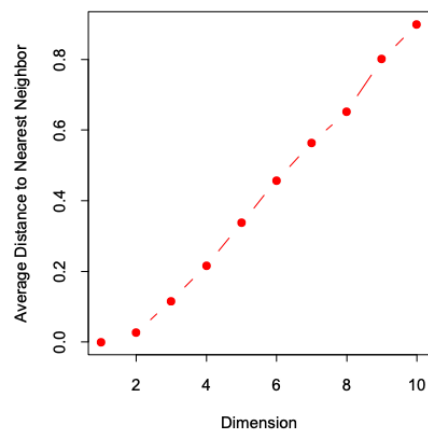
1-NN in One Dimension



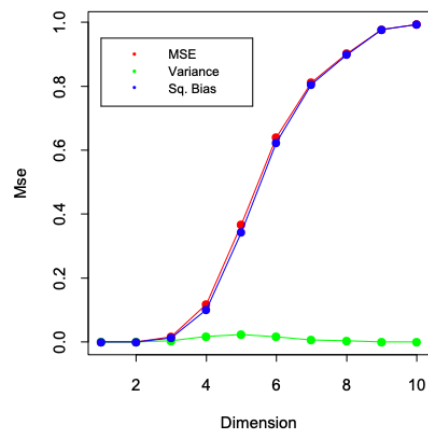
1-NN in One vs. Two Dimensions



Distance to 1-NN vs. Dimension



MSE vs. Dimension



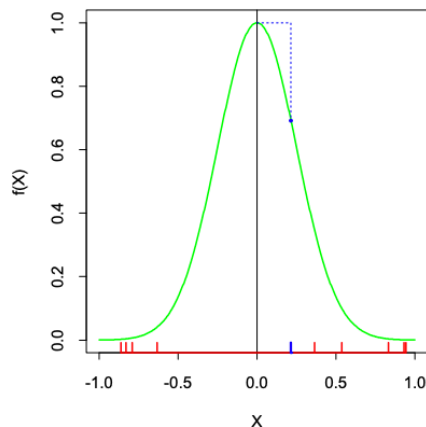
В чем состоит **проклятие размерности** для kNN?

- Расстояние до ближайшего соседа растет с ростом размерности.

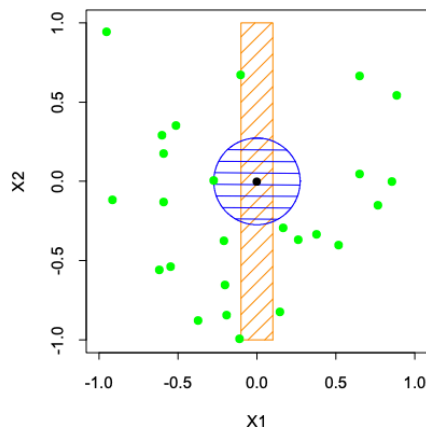


# Nearest Neighbors

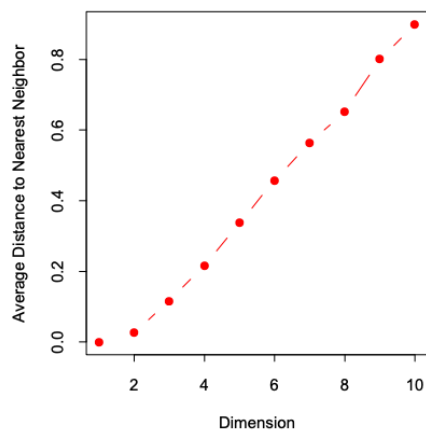
1-NN in One Dimension



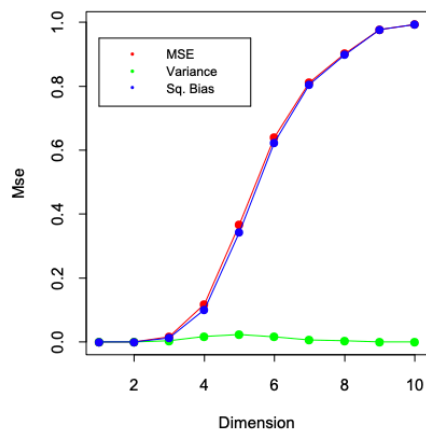
1-NN in One vs. Two Dimensions



Distance to 1-NN vs. Dimension



MSE vs. Dimension

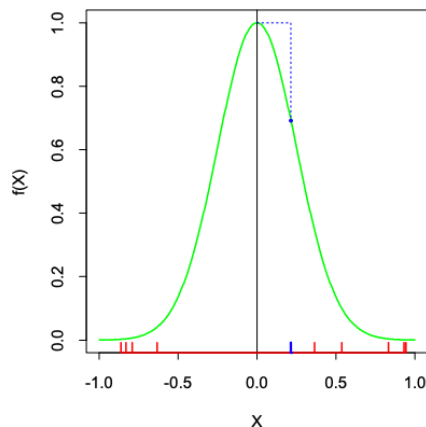


В чем состоит **проклятие размерности** для kNN?

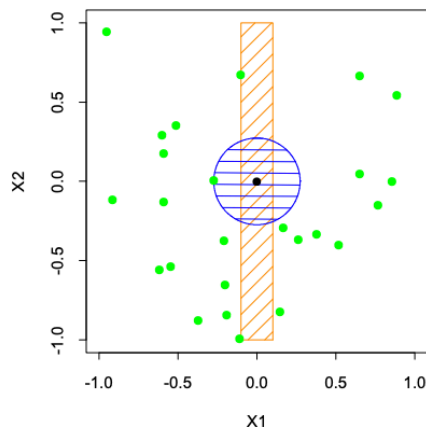
- Расстояние до ближайшего соседа растет с ростом размерности.
- По этой причине, растет bias нашей оценки.

# Nearest Neighbors

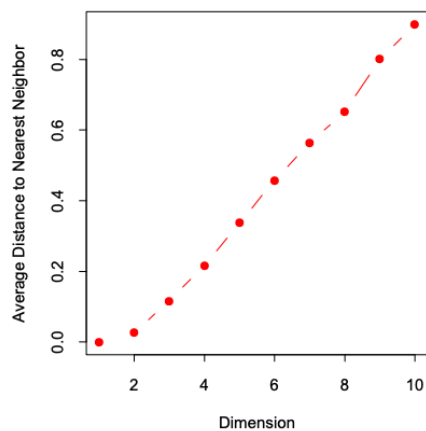
1-NN in One Dimension



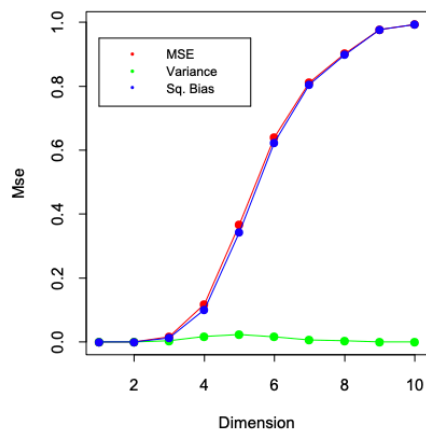
1-NN in One vs. Two Dimensions



Distance to 1-NN vs. Dimension



MSE vs. Dimension

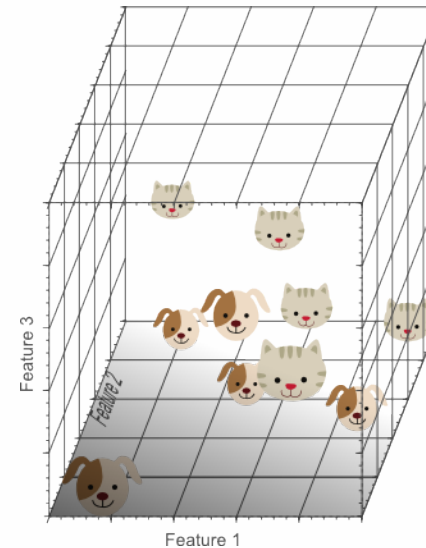
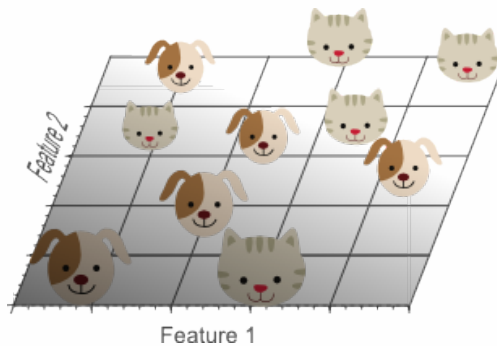


В чем состоит **проклятие размерности** для kNN?

- Расстояние до ближайшего соседа растет с ростом размерности.
- По этой причине, растет bias нашей оценки.
- Более того, с ростом размерности основная масса данных при определенных допущениях распределяется по поверхности гиперсферы, что делает ближайших соседей практически неотличимыми от дальних.

# Nearest Neighbors

Проклятие размерности  $\lim_{d \rightarrow \infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} \rightarrow 0$



# Naïve Bayes

## Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



Thomas Bayes  
1702 - 1761

Используется для классификации

- Преимущества:

- Скорость
- Малое число гиперпараметров
- Успешно работает в многомерных пространствах

# Naïve Bayes

## Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



Thomas Bayes  
1702 - 1761

Используется для классификации

### • Преимущества:

- Скорость
- Малое число гиперпараметров
- Успешно работает в многомерных пространствах

### • Недостатки:

- Наивность

# Naïve Bayes

Пусть у нас бинарная классификация.

1. Предполагаем распределение наших признаков (например, нормальное).

$$P(x_i) \sim N(m_i, \sigma_i)$$

2. Используя bayes rule записываем:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

здесь  $X = \{x_1, x_2, \dots, x_m\}$

# Naïve Bayes

Главное **допущение**, которое делается далее (**наивность**) – взаимная независимость признаков. То есть:

$$P(X) = P(x_1 \cdot \dots \cdot x_m) = P(x_1) \cdot \dots \cdot P(x_m)$$

# Naïve Bayes

Главное **допущение**, которое делается далее (**наивность**) – взаимная независимость признаков. То есть:

$$P(X) = P(x_1 \cdot \dots \cdot x_m) = P(x_1) \cdot \dots \cdot P(x_m)$$

Отсюда переходим к финальной проблеме:

$$\begin{aligned}\hat{y} &= \arg \max_y \frac{P(X|y)P(y)}{P(X)} = \\ &= \arg \max_y P(X|y)P(y) = \\ &= \arg \max_y P(y) \prod_i P(x_i|y)\end{aligned}$$



# Naïve Bayes

Далее для численной стабильности желательно перейти к логарифмам:

$$\hat{y} = \arg \max_y \{ \ln(P(y)) + \sum_i \ln(P(x_i|y)) \}$$

Bottomline:

Необходимо быть внимательным к распределениям переменных!

# Linear models. Постановка задачи.

Случай двух переменных:

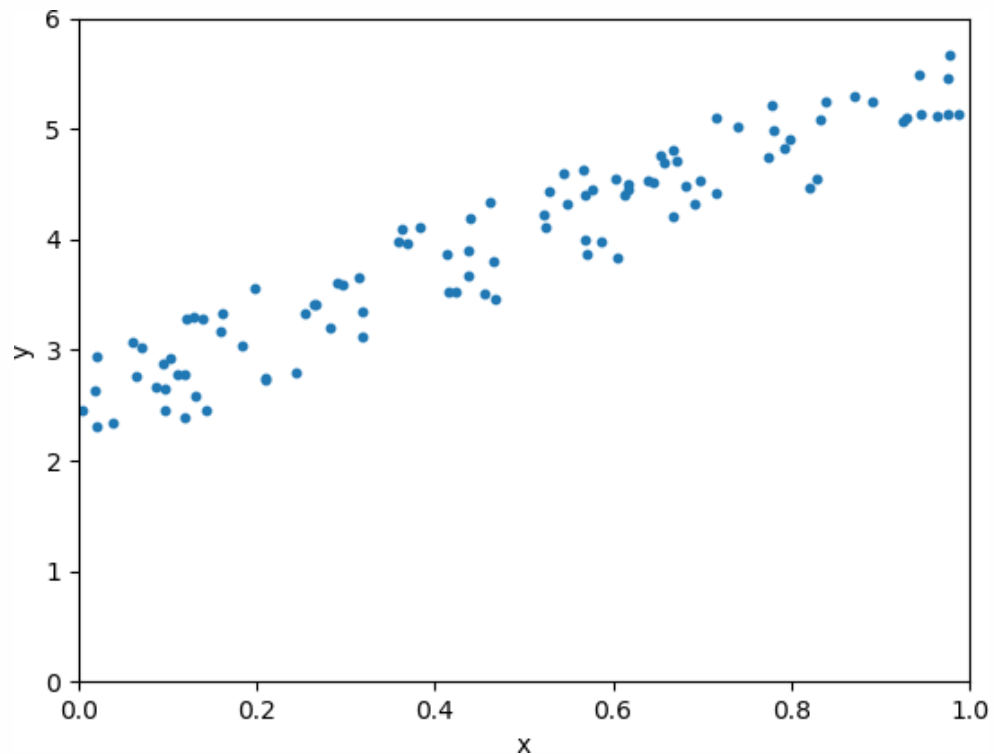


Рис.: Данные

# Linear models. Постановка задачи.

Случай двух переменных:

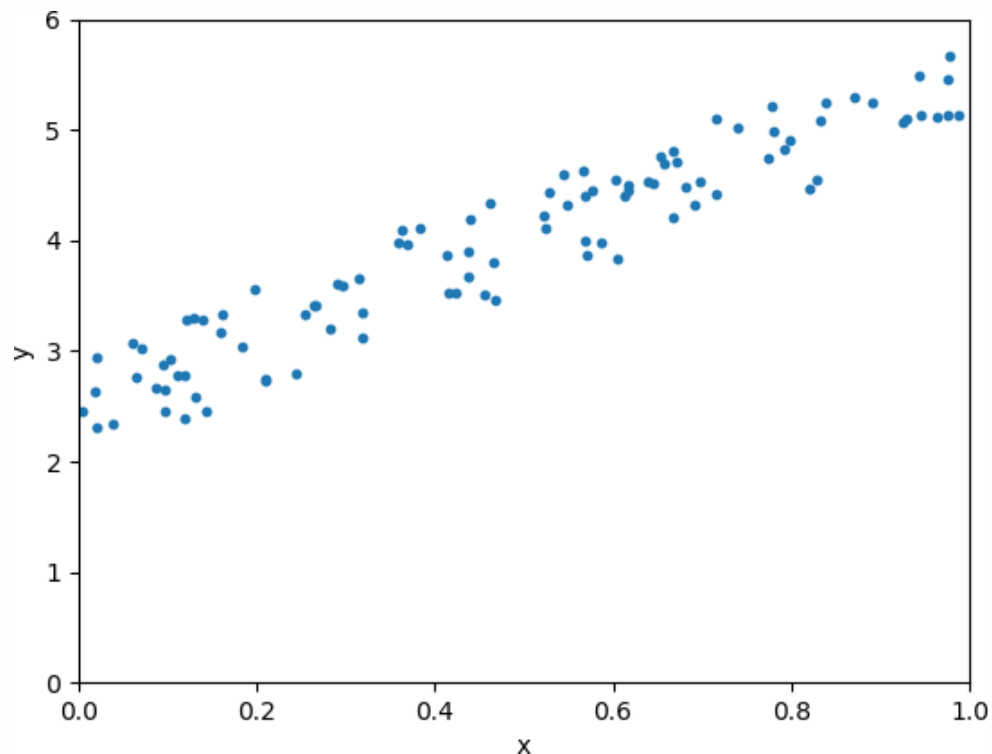


Рис.: Данные

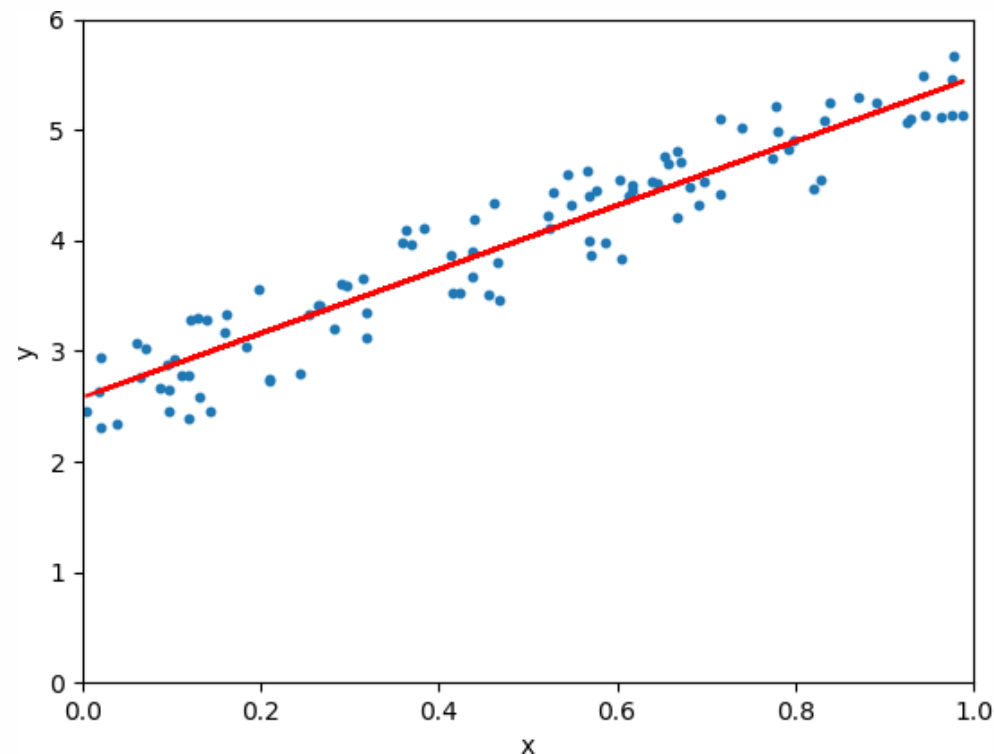


Рис.: Регрессия

Пусть модель распределения данных линейна по параметрам и имеет вид:

$$y = \beta_0 + \beta_1 x + \epsilon$$

# Linear models. Постановка задачи.

- Пусть есть одна целевая переменная  $y$  и  $k$  признаков  $x_1, \dots, x_k$ .  
Тогда линейная модель имеет вид:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

# Linear models. Постановка задачи.

- Пусть есть одна целевая переменная  $y$  и  $k$  признаков  $x_1, \dots, x_k$ .  
Тогда линейная модель имеет вид:

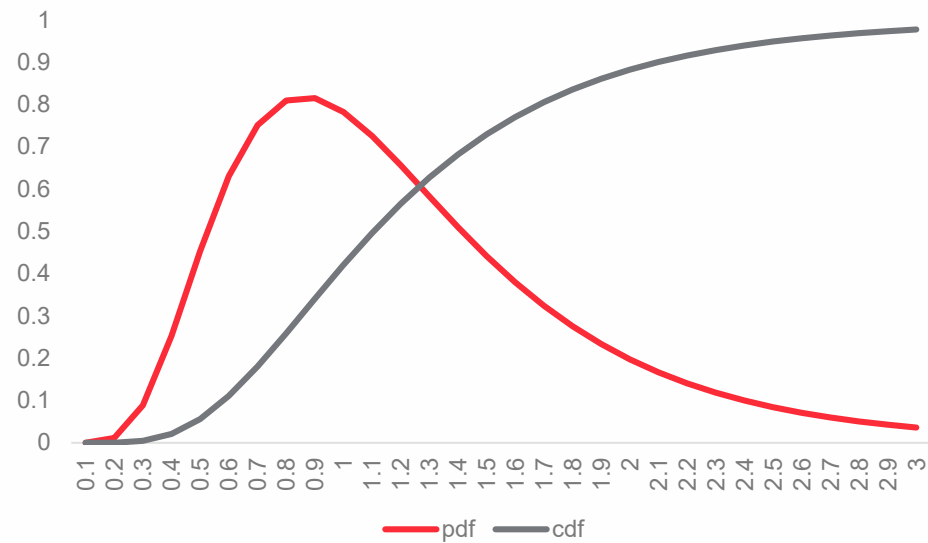
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

- Наша задача - получить оценку параметров модели  $\beta_i$   
В итоге мы хотим получить уравнение:

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

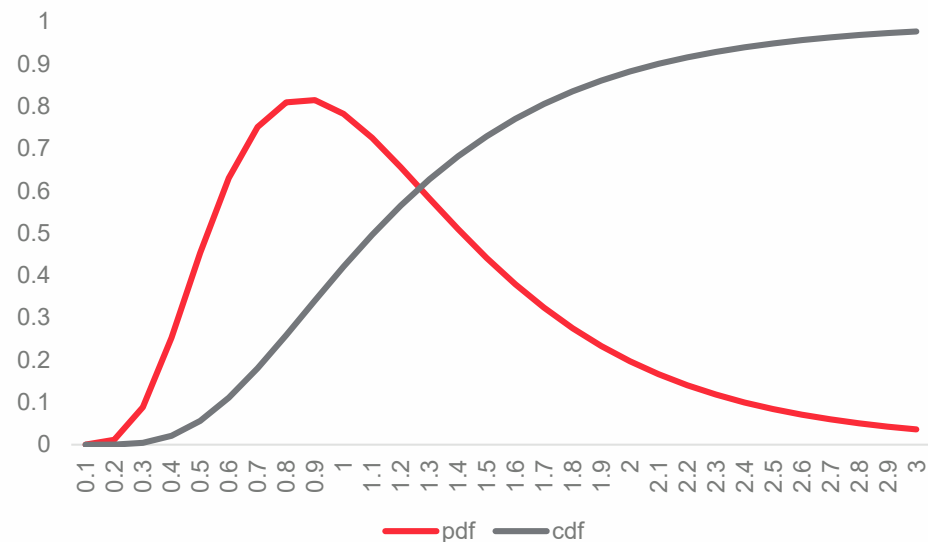
# Linear models. CEF.

Пусть имеется переменная – затраты абонента на связь.  
Есть распределение затрат по абонентам.



# Linear models. CEF.

Пусть имеется переменная – затраты абонента на связь.  
Есть распределение затрат по абонентам.



Один из параметров, интересующих нас – матожидание:

$$E(y) = \int_{-\infty}^{\infty} u f(u) du$$

# Linear models. CEF.

Рассмотрим влияние различных переменных на затраты (пол, возраст и т.д.). Тогда можем рассмотреть условное матожидание при различных значениях этих параметров:

$$E(y|sex = man, age = 24, education = higher, ...) = ...$$



# Linear models. CEF.

Рассмотрим влияние различных переменных на затраты (пол, возраст и т.д.). Тогда можем рассмотреть условное матожидание при различных значениях этих параметров:

$$E(y|sex = man, age = 24, education = higher, ...) = ...$$

Можем записать условное матожидание в общем виде:

$$E(y|x_1, x_2, ..., x_k) = m(x_1, x_2, ..., x_k) = m(\mathbf{x})$$

# Linear models. CEF.

Рассмотрим влияние различных переменных на затраты (пол, возраст и т.д.). Тогда можем рассмотреть условное матожидание при различных значениях этих параметров:

$$E(y|sex = man, age = 24, education = higher, ...) = ...$$

Можем записать условное матожидание в общем виде:

$$E(y|x_1, x_2, ..., x_k) = m(x_1, x_2, ..., x_k) = m(\mathbf{x})$$

$m(\mathbf{x})$  можно представить в виде:

$$m(\mathbf{x}) = E(y|\mathbf{x}) = \int_{-\infty}^{\infty} f(y|\mathbf{x}) dy$$



# Linear models. CEF error.

Определим понятие CEF error как:

$$e = y - m(\mathbf{x})$$

# Linear models. Prediction error.

- Введем понятие ошибки предсказания:

$$E((g(\mathbf{x}) - y)^2)$$

# Linear models. Prediction error.

- Введем понятие ошибки предсказания:

$$E((g(\mathbf{x}) - y)^2)$$

- Можно утверждать, что для любой  $g(x)$ :

$$E((g(\mathbf{x}) - y)^2) \geq E((m(x) - y)^2)$$

# Linear models. Prediction error.

- Введем понятие ошибки предсказания:

$$E((g(\mathbf{x}) - y)^2)$$

- Можно утверждать, что для любой  $g(x)$ :

$$E((g(\mathbf{x}) - y)^2) \geq E((m(x) - y)^2)$$

Таким образом,  $m(x)$  – MMSE предиктор, решающий задачу:

$$m(\mathbf{x}) = \arg \min_{g(x)} E((g(x) - y)^2)$$

# Linear models. OLS.

- Мы предположили, что данные генерируются линейной моделью:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \epsilon_i$$

# Linear models. OLS.

- Мы предположили, что данные генерируются линейной моделью:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \epsilon_i$$

- Предположим, также, что наблюдения  $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$  независимы и одинаково распределены (iid). Отсюда, предполагая линейность функции  $g(\mathbf{x})$ , запишем задачу:

$$\hat{\beta} = \arg \min_{\beta} E((y - \mathbf{x}\beta)^2)$$



# Linear models. OLS.

- Имея выборку из  $n$  наблюдений, перейдем к выборочной статистике для матожидания:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n ((y_i - \mathbf{x}_i \beta)^2)$$

# Linear models. OLS.

- Имея выборку из  $n$  наблюдений, перейдем к выборочной статистике для матожидания:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n ((y_i - \mathbf{x}_i \beta)^2)$$

- Из условий минимума первого порядка (FOC) имеем:

$$\sum_{i=1}^n \mathbf{x}'_i (y_i - \mathbf{x}_i \hat{\beta}) = 0$$

# Linear models. OLS.

- Имея выборку из  $n$  наблюдений, перейдем к выборочной статистике для матожидания:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n ((y_i - \mathbf{x}_i \beta)^2)$$

- Из условий минимума первого порядка (FOC) имеем:

$$\sum_{i=1}^n \mathbf{x}'_i (y_i - \mathbf{x}_i \hat{\beta}) = 0$$

- В матричном виде модель выглядит:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

# Linear models. OLS.

- Имея выборку из  $n$  наблюдений, перейдем к выборочной статистике для матожидания:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n ((y_i - \mathbf{x}_i \beta)^2)$$

- Из условий минимума первого порядка (FOC) имеем:

$$\sum_{i=1}^n \mathbf{x}'_i (y_i - \mathbf{x}_i \hat{\beta}) = 0$$

- В матричном виде модель выглядит:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

- И FOC:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$

## Linear models. OLS.

В итоге получаем решение:

i

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

!

# Linear models. Assumptions OLS.

Assumption 1 (экзогенность):

$$E(\mathbf{x}'\epsilon) = \mathbf{0}$$

Это предположение о том, что ошибка имеет нулевое среднее и некоррелирована с регрессорами.

# Linear models. Assumptions OLS.

Assumption 1 (экзогенность):

$$E(\mathbf{x}'\epsilon) = \mathbf{0}$$

Это предположение о том, что ошибка имеет нулевое среднее и некоррелирована с регрессорами.

Assumption 2 (полный ранг):

$$\text{rank}(\mathbf{x}'\mathbf{x}) = K$$

Эквивалентно предположению о положительно определенной матрице.

# Linear models. Assumptions OLS.

Assumption 1 (экзогенность):

$$E(\mathbf{x}'\epsilon) = \mathbf{0}$$

Это предположение о том, что ошибка имеет нулевое среднее и некоррелирована с регрессорами.

Assumption 2 (полный ранг):

$$\text{rank}(\mathbf{x}'\mathbf{x}) = K$$

Эквивалентно предположению о положительно определенной матрице.

Теорема (состоятельность МНК):

При выполнении предположений 1 и 2 оценка  $\hat{\beta}$  является состоятельной оценкой параметра  $\beta$  из модели  $y = \beta\mathbf{x} + \epsilon$



# Linear models. Assumptions OLS.

Assumption 3 (гомоскедастичность):

$$E(\epsilon^2 \mathbf{x}' \mathbf{x}) = \sigma^2 E(\mathbf{x}' \mathbf{x})$$

Где  $\sigma^2 = E(\epsilon^2)$ .

Квадрат ошибки некоррелирован с каждым элементом, их квадратами и их кросс-продуктам.

Из свойств условных матожиданий видно, что достаточным условием является  $var(\epsilon|\mathbf{x}) = \sigma^2$

# Linear models. Assumptions OLS.

Assumption 3 (гомоскедастичность):

$$E(\epsilon^2 \mathbf{x}' \mathbf{x}) = \sigma^2 E(\mathbf{x}' \mathbf{x})$$

Где  $\sigma^2 = E(\epsilon^2)$ .

Квадрат ошибки некоррелирован с каждым элементом, их квадратами и их кросс-продуктам.

Из свойств условных матожиданий видно, что достаточным условием является  $var(\epsilon|\mathbf{x}) = \sigma^2$

Теорема (асимптотическая нормальность МНК):

При выполнении предположений 1, 2 и 3:

$$\sqrt{N}(\hat{\beta} - \beta) \sim \mathcal{N}(0, V_{\beta})$$

$$\sqrt{N}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \sigma^2 E(\mathbf{x}' \mathbf{x})^{-1})$$

# Linear models. Нормальная регрессия.

○ Предположим, что **ошибка распределена нормально**. Тогда мы имеем модель:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon,$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

# Linear models. Нормальная регрессия.

- Предположим, что **ошибка распределена нормально**. Тогда мы имеем модель:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon,$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

- **Правдоподобием** называют совместную вероятность реализовавшейся выборки, рассматривая её как функцию от параметров. Для модели выше:

$$f(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}'\beta)^2\right)$$

# Linear models. ММП.

Записывая условную плотность для всей выборки, получим функцию правдоподобия:

$$f(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n f(y_i | \mathbf{x}_i)$$

# Linear models. ММП.

Записывая условную плотность для всей выборки, получим функцию правдоподобия:

$$f(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n f(y_i | \mathbf{x}_i)$$

Логарифмируя и находя максимум из FOC:

$$(\hat{\beta}_{mle}, \hat{\sigma}_{mle}^2) = \arg \max_{\beta \in \mathbb{R}, \sigma^2 > 0} \ln(L(\beta, \sigma^2))$$



# Linear models. t-statistics.

Рассмотрим следующую статистику:

$$T(\beta) = \frac{\hat{\beta} - \beta}{\sqrt{V_{\hat{\beta}}}}$$

# Linear models. t-statistics.

Рассмотрим следующую статистику:

$$T(\beta) = \frac{\hat{\beta} - \beta}{\sqrt{V_{\hat{\beta}}}}$$

Можем получить:

$$T(\beta) = \frac{\hat{\beta} - \beta}{\sqrt{V_{\hat{\beta}}}} = \frac{\sqrt{N}(\hat{\beta} - \beta)}{\sqrt{V_{\beta}}} \xrightarrow{d} \frac{\mathcal{N}(0, V_{\beta})}{\sqrt{V_{\beta}}} = Z \sim \mathcal{N}(0, 1)$$




# Linear models. t-statistics.

Так как вместо  $V_{\hat{\beta}}$  мы имеем лишь её оценку  $\hat{V}_{\hat{\beta}}$ , то после некоторых преобразований можем получить:

$$T(\beta) = t = \frac{\hat{\beta} - \beta}{\sqrt{\hat{V}_{\hat{\beta}}}} \sim \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi_{n-k}^2}{n-k}}} \sim t_{n-k}$$



# Linear models. t-statistics.



Таким образом, можем проверить значимость коэффициентов регрессии. Проверяем гипотезу  
 $H_0 : \beta = 0$

# Linear models. t-statistics.

Таким образом, можем проверить значимость коэффициентов регрессии. Проверяем гипотезу  $H_0 : \beta = 0$

Считаем статистику:

$$t = \frac{\hat{\beta}}{\sqrt{\hat{V}_{\hat{\beta}}}} \sim t_{n-k}$$

при справедливости  $H_0$ .

# Linear models. t-statistics.

Таким образом, можем проверить значимость коэффициентов регрессии. Проверяем гипотезу  $H_0 : \beta = 0$

Считаем статистику:

$$t = \frac{\hat{\beta}}{\sqrt{\hat{V}_{\hat{\beta}}}} \sim t_{n-k}$$

при справедливости  $H_0$ .

Её **p-value** – это  $P(\text{reject } H_0 | H_0)$ . Уровень значимости – некоторая установленная нами граница такая, что, если p-value меньше либо равно данной границы, мы отвергаем гипотезу  $H_0$  на этом уровне значимости.



# Linear models. F-test.

Предположим, мы хотим протестировать множественную гипотезу:

$$H_0 : \beta_0 = 0, \dots, \beta_k = 0$$

в этом нам поможет F-test.

# Linear models. F-test.

Предположим, мы хотим протестировать множественную гипотезу:

$$H_0 : \beta_0 = 0, \dots, \beta_k = 0$$

в этом нам поможет F-test.

F статистика выглядит следующим образом:

$$F = \frac{\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{q}}{\frac{\hat{\sigma}^2}{n-k}}$$

# Linear models. F-test.

Предположим, мы хотим протестировать множественную гипотезу:

$$H_0 : \beta_0 = 0, \dots, \beta_k = 0$$

в этом нам поможет F-test.

F статистика выглядит следующим образом:

$$F = \frac{\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{q}}{\frac{\hat{\sigma}^2}{n-k}}$$

где

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^n (y_i - \mathbf{x}_i \hat{\beta})^2$$

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{i=1}^n (y_i - \mathbf{x}_i \tilde{\beta})^2$$

# Linear models. F-test.

$\tilde{\beta}, \tilde{\sigma}$  - это оценки restricted модели, где соответствующие коэффициенты приняты равными 0.

$q = df_r - df_{ur}$ , где степень свободы определяется как число наблюдений минус число параметров.

F статистика имеет распределение:  $F \sim F_{q, n-k-1}$



# Linear models. Градиентный спуск.

В случае, когда мы имеем дело с большими данными, иногда невозможно применить МНК (хотя существуют реализации и для этих случаев).

Применяют метод градиентного спуска. Минимизируем:

$$L(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2$$

# Linear models. Градиентный спуск.

В случае, когда мы имеем дело с большими данными, иногда невозможно применить МНК (хотя существуют реализации и для этих случаев).

Применяют метод градиентного спуска. Минимизируем:

$$L(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2$$

Обновляем значение:

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} - \alpha \nabla L(\hat{\beta}^{(i)})$$

# Linear models. Градиентный спуск.

В случае, когда мы имеем дело с большими данными, иногда невозможно применить МНК (хотя существуют реализации и для этих случаев).

Применяют метод градиентного спуска. Минимизируем:

$$L(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}_i \beta)^2$$

Обновляем значение:

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} - \alpha \nabla L(\hat{\beta}^{(i)})$$

Останавливаемся при

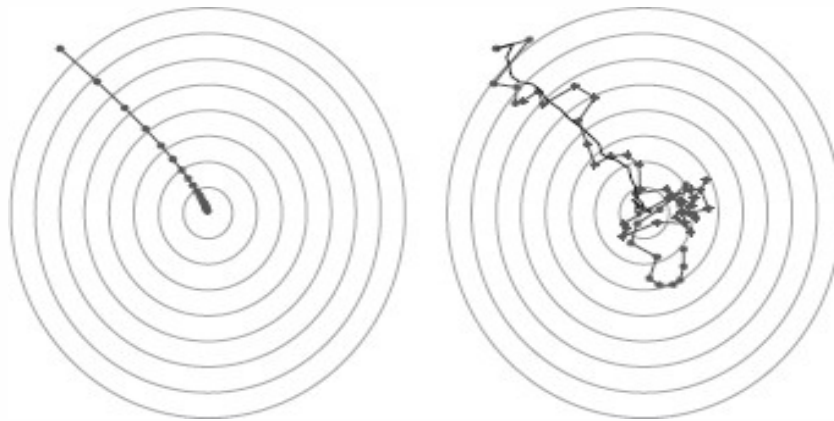
$$|\hat{\beta}^{(i+1)} - \hat{\beta}^{(i)}| < \epsilon$$

# Linear models. Стохастический градиентный спуск.

При вычислении градиента на каждой итерации суммируем значения. Долго.


Альтернатива – стохастический градиентный спуск. Считаем градиент по каждому наблюдению.

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} - \alpha \nabla L_i(\hat{\beta}^{(i)})$$





# Linear models. Classification.



Решаем задачу классификации. Имеем  $y \in \{1, 2, \dots, K\}$

Как будем решать?



# Linear models. Classification.

Решаем задачу классификации. Имеем  $y \in \{1, 2, \dots, K\}$

Как будем решать?

- One vs. One



# Linear models. Classification.

Решаем задачу классификации. Имеем  $y \in \{1, 2, \dots, K\}$

Как будем решать?

- One vs. One
- One vs. All



# Linear models. Classification.

Решаем задачу классификации. Имеем  $y \in \{1, 2, \dots, K\}$

Как будем решать?

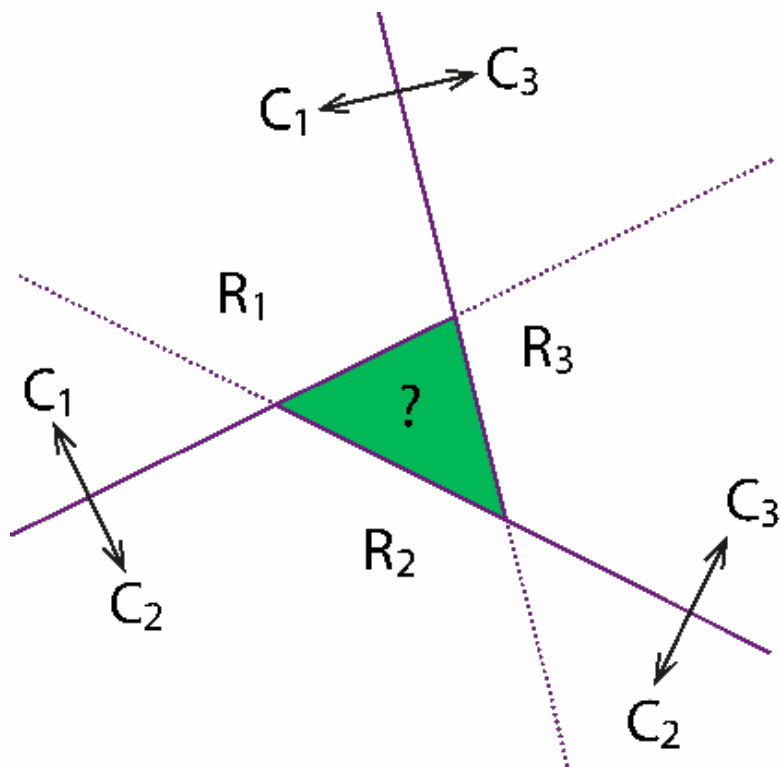
- One vs. One
- One vs. All
- Multiclass



# Linear models. One vs. One.

Строим  $K(K-1)/2$  бинарных классификаторов.

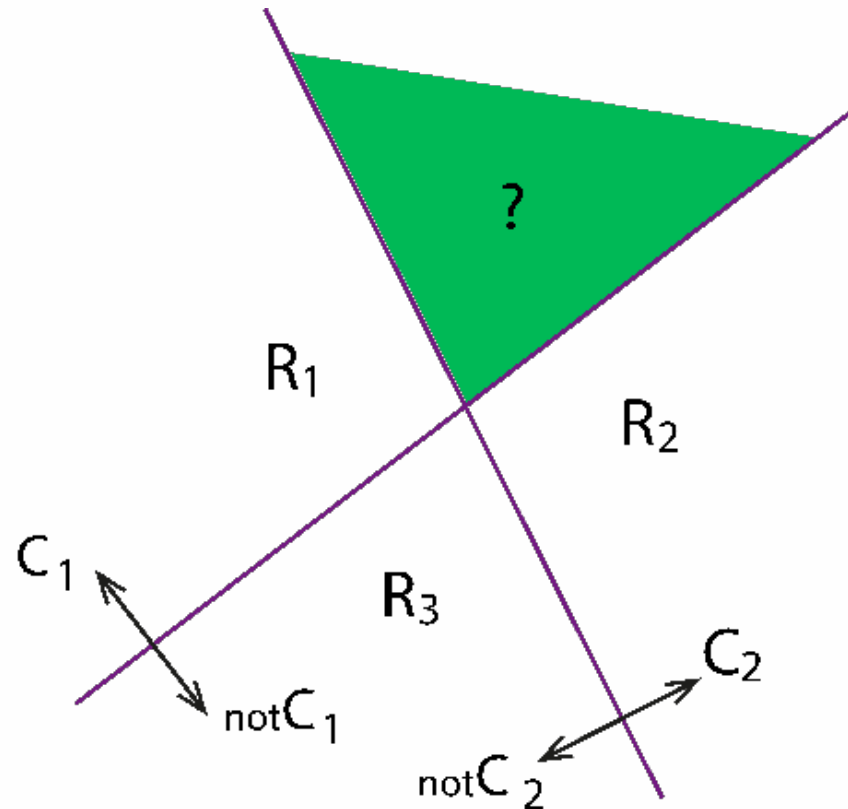
Проблема:



# Linear models. One vs. All.


Строим  $K-1$  бинарных классификаторов.

Проблема:





# Linear models. Logistic regression.



Предположим, что есть абонент, который выбирает купить или не купить услугу. Он покупает услугу, если "польза" от услуги больше нуля. Запишем:

$$y_i^* = \mathbf{x}_i' \mathbf{w} + \epsilon_i$$

# Linear models. Logistic regression.

Предположим, что есть абонент, который выбирает купить или не купить услугу.

Он покупает услугу, если "польза" от услуги больше нуля. Запишем:

$$\mathbf{y}_i^* = \mathbf{x}_i' \mathbf{w} + \epsilon_i$$

Модель скрытой переменной, где  $\mathbf{y}_i^*$  - ненаблюдаемая переменная, а  $\epsilon_i$  - ошибка, имеющая распределение F.

$$\mathbf{y}_i = \begin{cases} 1, & \text{if } \mathbf{y}_i^* \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

# Linear models. Logistic regression.

Предположим, что есть абонент, который выбирает купить или не купить услугу.

Он покупает услугу, если "польза" от услуги больше нуля. Запишем:

$$\mathbf{y}_i^* = \mathbf{x}_i' \mathbf{w} + \epsilon_i$$

Модель скрытой переменной, где  $\mathbf{y}_i^*$  - ненаблюдаемая переменная, а  $\epsilon_i$  - ошибка, имеющая распределение F.

$$\mathbf{y}_i = \begin{cases} 1, & \text{if } \mathbf{y}_i^* \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Тогда:

$$\begin{aligned} P(\mathbf{y}_i = 1 | \mathbf{x}_i) &= P(\mathbf{y}_i^* > 0 | \mathbf{x}_i) = P(\mathbf{x}_i' \mathbf{w} + \epsilon_i | \mathbf{x}_i) = \\ &= P(\epsilon_i > -\mathbf{x}_i' \mathbf{w} | \mathbf{x}_i) = 1 - F(-\mathbf{x}_i' \mathbf{w}) = F(\mathbf{x}_i' \mathbf{w}) \end{aligned}$$



# Linear models. Logistic regression.

Предположим, распределение ошибки – логистическое. То есть:

$$F(\epsilon) = \frac{1}{1 + \exp(-\epsilon)}$$

Отсюда получаем логистическую регрессию.

# Linear models. Logistic regression.

Можем подойти с другой стороны. Предположим, у нас 2 класса. Тогда:

$$p(y = 1|\mathbf{x}) = \frac{p(\mathbf{x}|y = 1)p(y = 1)}{p(\mathbf{x}|y = 0)p(y = 0) + p(\mathbf{x}|y = 1)p(y = 1)} = \frac{1}{1 + \exp(-\epsilon)} = \sigma(\epsilon)$$

# Linear models. Logistic regression.

Можем подойти с другой стороны. Предположим, у нас 2 класса. Тогда:

$$p(y = 1|\mathbf{x}) = \frac{p(\mathbf{x}|y = 1)p(y = 1)}{p(\mathbf{x}|y = 0)p(y = 0) + p(\mathbf{x}|y = 1)p(y = 1)} = \frac{1}{1 + \exp(-\epsilon)} = \sigma(\epsilon)$$

где  $\epsilon = \ln \frac{p(\mathbf{x}|y = 1)p(y = 1)}{p(\mathbf{x}|y = 0)p(y = 0)} = \ln \frac{p(y = 1|\mathbf{x})}{p(y = 0|\mathbf{x})}$  - отношение шансов.



# Linear models. Logistic regression.

Можем подойти с другой стороны. Предположим, у нас 2 класса. Тогда:

$$p(y = 1|\mathbf{x}) = \frac{p(\mathbf{x}|y = 1)p(y = 1)}{p(\mathbf{x}|y = 0)p(y = 0) + p(\mathbf{x}|y = 1)p(y = 1)} = \frac{1}{1 + \exp(-\epsilon)} = \sigma(\epsilon)$$

где  $\epsilon = \ln \frac{p(\mathbf{x}|y = 1)p(y = 1)}{p(\mathbf{x}|y = 0)p(y = 0)} = \ln \frac{p(y = 1|\mathbf{x})}{p(y = 0|\mathbf{x})}$  - отношение шансов.

При этом можем записать  $\epsilon = \ln \frac{\sigma}{1 - \sigma}$ , которую называют логит функцией.

# Linear models. Logistic regression.

Оценку коэффициентов получаем **методом максимального правдоподобия**.

Так как имеем **условные испытания бернулли**, запишем вероятность события:

$$f(y_i | \mathbf{x}_i) = p_i^{y_i} (1 - p_i)^{1-y_i} = F(\mathbf{x}_i' \mathbf{w})^{y_i} (1 - F(\mathbf{x}_i' \mathbf{w}))^{1-y_i}$$

Записываем функцию правдоподобия, логарифмируем её и находим максимум по коэффициентам из FOC. Получаем  $\hat{\mathbf{w}}$ .



# Linear models. Multiclass.

Рассмотрим multiclass logistic regression.

Предположим  $y \in \{1, 2, \dots, K\}$ .

# Linear models. Multiclass.

Рассмотрим multiclass logistic regression.

Предположим  $y \in \{1, 2, \dots, K\}$ .

Тогда, по аналогии с предыдущим запишем модель:

$$\ln \frac{P(y = 1|\mathbf{x})}{P(y = K|\mathbf{x})} = w_{10} + w'_1 \mathbf{x}$$

# Linear models. Multiclass.

Рассмотрим **multiclass logistic regression**.

Предположим  $y \in \{1, 2, \dots, K\}$ .

Тогда, по аналогии с предыдущим запишем модель:

$$\ln \frac{P(y = 1|\mathbf{x})}{P(y = K|\mathbf{x})} = w_{10} + w'_1 \mathbf{x}$$

$$\ln \frac{P(y = 2|\mathbf{x})}{P(y = K|\mathbf{x})} = w_{20} + w'_2 \mathbf{x}$$

...

$$\ln \frac{P(y = K - 1|\mathbf{x})}{P(y = K|\mathbf{x})} = w_{(K-1)0} + w'_{K-1} \mathbf{x}$$



# Linear models. Multiclass.

Отсюда легко получить:

$$P(y = i|\mathbf{x}) = \exp(w_{i0} + w'_i\mathbf{x})P(y = K|\mathbf{x})$$

# Linear models. Multiclass.

Отсюда легко получить:

$$P(y = i|\mathbf{x}) = \exp(w_{i0} + w'_i\mathbf{x})P(y = K|\mathbf{x})$$

$$P(y = K|\mathbf{x}) = 1 - \sum_{i=1}^{K-1} \exp(w_{i0} + w'_i\mathbf{x})P(y = K|\mathbf{x})$$

# Linear models. Multiclass.

Отсюда легко получить:

$$P(y = i|\mathbf{x}) = \exp(w_{i0} + w'_i\mathbf{x})P(y = K|\mathbf{x})$$

$$P(y = K|\mathbf{x}) = 1 - \sum_{i=1}^{K-1} \exp(w_{i0} + w'_i\mathbf{x})P(y = K|\mathbf{x})$$

$$P(y = K|\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(w_{i0} + w'_i\mathbf{x})}$$



# Linear models. Multiclass.

Отсюда легко получить:

$$P(y = i|\mathbf{x}) = \exp(w_{i0} + w'_i\mathbf{x})P(y = K|\mathbf{x})$$

$$P(y = K|\mathbf{x}) = 1 - \sum_{i=1}^{K-1} \exp(w_{i0} + w'_i\mathbf{x})P(y = K|\mathbf{x})$$

$$P(y = K|\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(w_{i0} + w'_i\mathbf{x})}$$

Далее действуем, применяя метод максимального правдоподобия и получаем оценки для параметров  $w$ .

# Linear models. Bias-variance decomposition.

Рассмотрим разложение MSE следующим образом:

$$\begin{aligned}MSE &= E(y - \hat{m}(\mathbf{x}))^2 = \sigma_\epsilon^2 + (Em(\mathbf{x}) - \hat{m}(\mathbf{x}))^2 + E(m(\mathbf{x}) - Em(\mathbf{x}))^2 = \\&= \sigma_\epsilon^2 + Bias^2(\hat{m}(\mathbf{x})) + var(\hat{m}(\mathbf{x}))\end{aligned}$$

# Linear models. Bias-variance decomposition.

Рассмотрим разложение MSE следующим образом:

$$\begin{aligned}MSE &= E(y - \hat{m}(\mathbf{x}))^2 = \sigma_\epsilon^2 + (Em(\mathbf{x}) - \hat{m}(\mathbf{x}))^2 + E(m(\mathbf{x}) - Em(\mathbf{x}))^2 = \\&= \sigma_\epsilon^2 + Bias^2(\hat{m}(\mathbf{x})) + var(\hat{m}(\mathbf{x}))\end{aligned}$$

Ошибка раскладывается на **неустранимую ошибку, bias и variance**.

# Linear models. Bias-variance decomposition.

Рассмотрим разложение MSE следующим образом:

$$\begin{aligned}MSE &= E(y - \hat{m}(\mathbf{x}))^2 = \sigma_\epsilon^2 + (Em(\mathbf{x}) - \hat{m}(\mathbf{x}))^2 + E(m(\mathbf{x}) - Em(\mathbf{x}))^2 = \\&= \sigma_\epsilon^2 + Bias^2(\hat{m}(\mathbf{x})) + var(\hat{m}(\mathbf{x}))\end{aligned}$$

Ошибка раскладывается на **неустранимую ошибку, bias и variance**.

При добавлении новых компонент в  $\mathbf{x}$  мы уменьшаем **bias**, но увеличиваем **дисперсию**.

# Linear models. Bias-variance decomposition.

Рассмотрим разложение MSE следующим образом:

$$\begin{aligned}MSE &= E(y - \hat{m}(\mathbf{x}))^2 = \sigma_\epsilon^2 + (Em(\mathbf{x}) - \hat{m}(\mathbf{x}))^2 + E(m(\mathbf{x}) - Em(\mathbf{x}))^2 = \\&= \sigma_\epsilon^2 + Bias^2(\hat{m}(\mathbf{x})) + var(\hat{m}(\mathbf{x}))\end{aligned}$$

Ошибка раскладывается на **неустранимую ошибку, bias и variance**.

При добавлении новых компонент в  $\mathbf{x}$  мы уменьшаем **bias**, но увеличиваем **дисперсию**.

Вопрос – как подобрать  $\mathbf{x}$  оптимально?



# Linear models. Stepwise selection.

Отличают **forward** и **backward stepwise selection**.

**Forward stepwise selection** – жадный алгоритм добавления признаков.

# Linear models. Stepwise selection.

Отличают **forward** и **backward stepwise selection**.

**Forward stepwise selection** – жадный алгоритм добавления признаков.

Плюсы:

- Можно применять при  $k \gg n$
- Имеет меньшую дисперсию, но, возможно, большее смещение

# Linear models. Stepwise selection.

Отличают **forward** и **backward stepwise selection**.

**Forward stepwise selection** – жадный алгоритм добавления признаков.

Плюсы:

- Можно применять при  $k \gg n$
- Имеет меньшую дисперсию, но, возможно, большее смещение

**Backward** – начинается с полной модели и убирает признаки по одному.





# Linear models. Regularization.

Какие проблемы может решать регуляризация?



# Linear models. Regularization.

Какие проблемы может решать регуляризация?

- Overfitting
- Feature selection
- High variance



# Linear models. Regularization.

Какие проблемы может решать регуляризация?

- Overfitting
- Feature selection
- High variance

Популярны два основных типа регуляризации – Ridge(L2) и Lasso(L1).

# Linear models. Ridge regression.

Решает проблему больших по модулю коэффициентов для коррелированных переменных, производя “weight decay”.

$$\hat{\beta}_{ridge} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^k \beta_j^2 \right\}$$

# Linear models. Ridge regression.

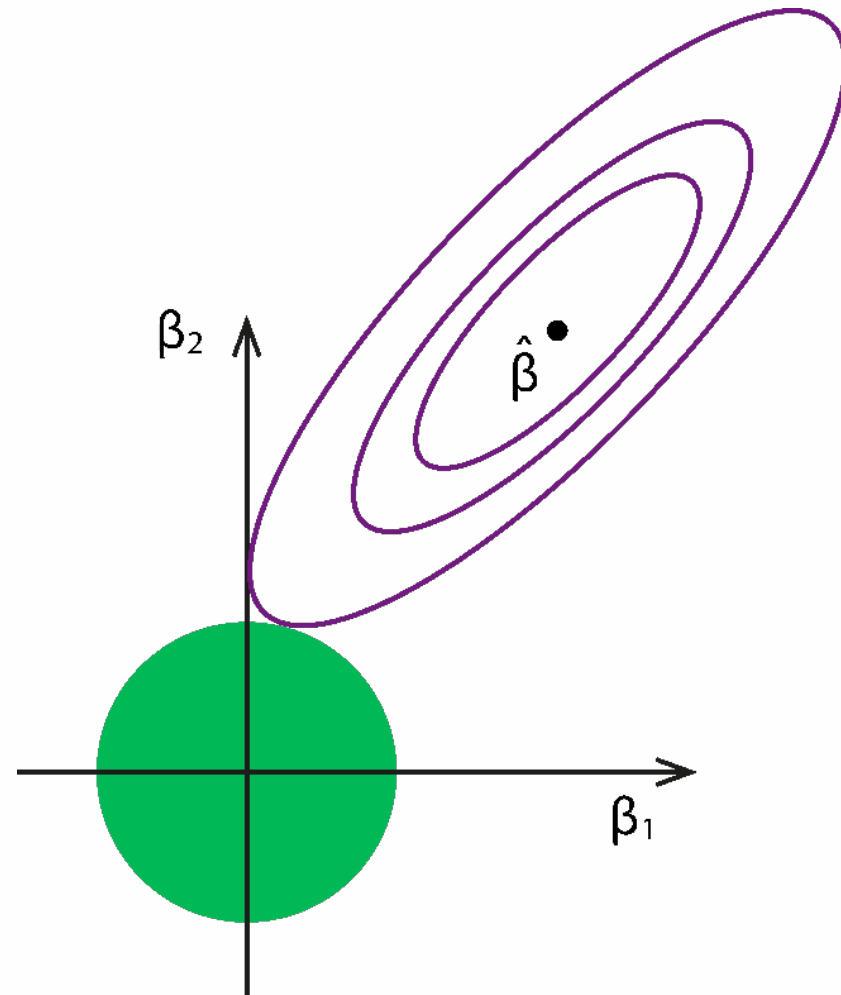
Решает проблему больших по модулю коэффициентов для коррелированных переменных, производя “weight decay”.

$$\hat{\beta}_{ridge} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^k \beta_j^2 \right\}$$

Эквивалентно задаче условной минимизации:

$$\begin{aligned} \hat{\beta}_{ridge} = \arg \min_{\beta} & \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k x_{ij} \beta_j)^2 \\ \text{s.t.} & \sum_{j=1}^k \beta_j^2 \leq t \end{aligned}$$

# Linear models. Ridge regression.



# Linear models. Lasso regression.

В дополнение к проблемам ridge решает проблему отбора признаков.

$$\hat{\beta}_{lasso} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^k |\beta_j| \right\}$$

# Linear models. Lasso regression.

В дополнение к проблемам ridge решает проблему отбора признаков.

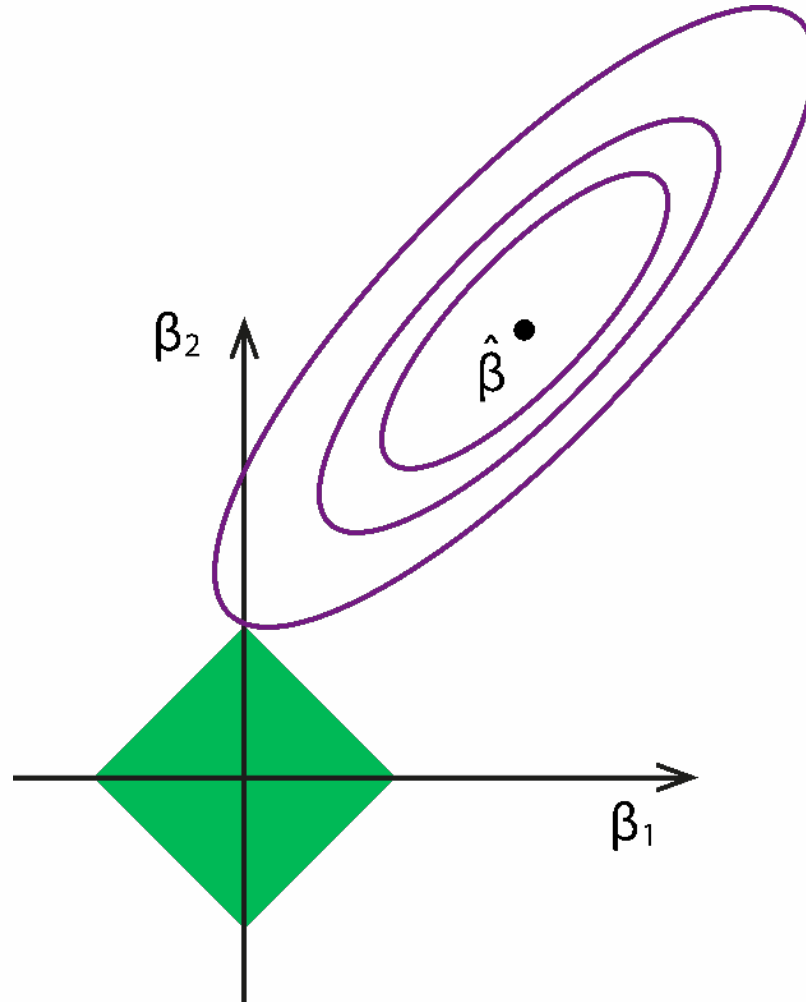
$$\hat{\beta}_{lasso} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^k |\beta_j| \right\}$$

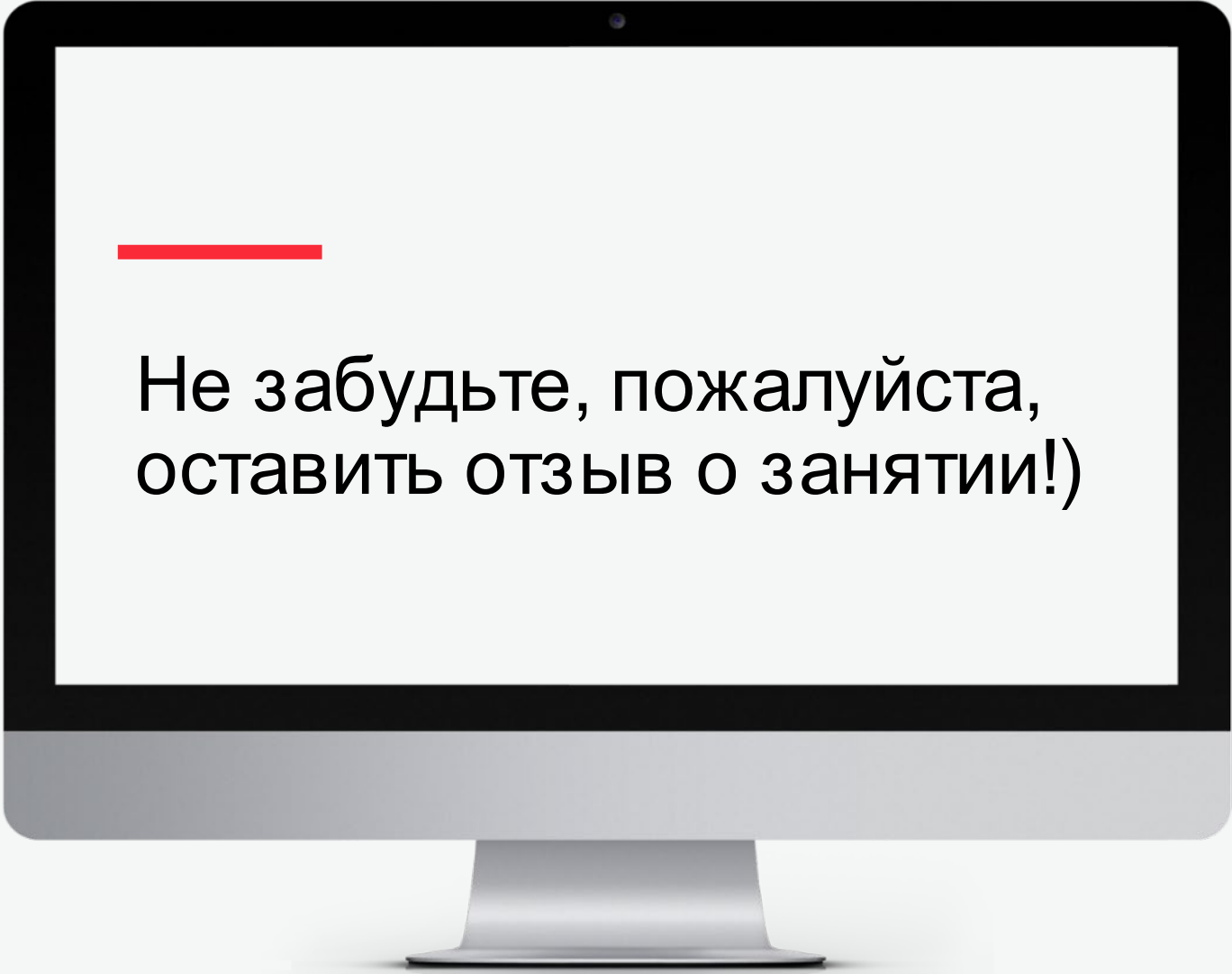
Эта задача эквивалентна задаче условной минимизации:

$$\begin{aligned} \hat{\beta}_{lasso} = \arg \min_{\beta} & \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k x_{ij} \beta_j)^2 \\ \text{s.t.} & \sum_{j=1}^k |\beta_j| \leq t \end{aligned}$$

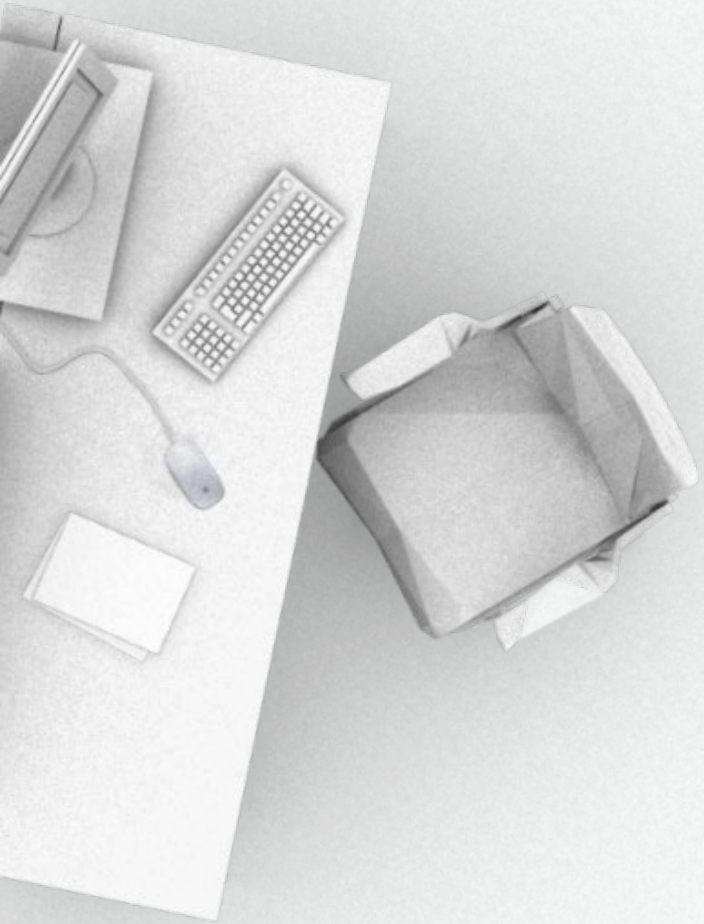


# Linear models. Lasso regression.



A computer monitor with a black bezel and a silver base. The screen is white and displays a red horizontal line followed by the text "Не забудьте, пожалуйста, оставить отзыв о занятии!".

—  
Не забудьте, пожалуйста,  
оставить отзыв о занятии!)



# Задачи классификации и регрессии

Практическая часть