Основные задачи ML
Базовые методы
Линейные модели

Лекция 2

Иван Горбань

План занятия:

1. Виды моделей
   a. Категоризация по виду модели
   b. Категоризация по задаче
   c. Категоризация по данным
2. Bias-Variance tradeoff и train-test
3. Методы
   a. kNN
   b. Наивный Байес
   c. Линейные модели
4. Подробнее о линейных моделях
   a. Линейная регрессия
   b. Логистическая регрессия
   c. Регуляризация

Виды моделей

Категоризации:

Виды моделей

Категоризации:

Виды моделей

Категоризации:

I. Область применения
   A. Объясняющие (SL)
   B. Предиктивные (ML)

Виды моделей

Категоризации:

I. Область применения
   A. Объясняющие (SL)
   B. Предиктивные (ML)

II. Задача:
   A. Обучение с учителем

Виды моделей

Категоризации:

I. Область применения
   A. Объясняющие (SL)
   B. Предиктивные (ML)
II. Задача:
   A. Обучение с учителем
      a. Задача регрессии
      b. Задача классификации

Виды моделей

Категоризации:

I. Область применения
   A. Объясняющие (SL)
   B. Предиктивные (ML)
II. Задача:
   A. Обучение с учителем
      a. Задача регрессии
      b. Задача классификации
   B. Обучение без учителя

Виды моделей

Категоризации:

I. Область применения
   A. Объясняющие (SL)
   B. Предиктивные (ML)

II. Задача:
   A. Обучение с учителем
      a. Задача регрессии
      b. Задача классификации
   B. Обучение без учителя

III. Характер данных:
   A. Cross-Section
   B. Time Series
   C. Panel

Предиктивные vs Объясняющие

Предиктивные vs Объясняющие

Объясняющие - легко интерпретируемые модели → инсайты

Предиктивные vs Объясняющие

Объясняющие - легко интерпретируемые модели → инсайты

Основная задача – понять закономерности происходящих процессов, найдя зависимости в данных.

Предиктивные vs Объясняющие

Объясняющие - легко интерпретируемые модели → инсайты

Основная задача – понять закономерности происходящих процессов, найдя зависимости в данных.

Акцент - на статистической значимости результатов.

Предиктивные vs Объясняющие

Объясняющие - легко интерпретируемые модели → инсайты

Основная задача – понять закономерности происходящих процессов, найдя зависимости в данных.

Акцент - на статистической значимости результатов.

Примеры: линейные модели, деревья решений, k-means кластеризация

Предиктивные vs Объясняющие

Объясняющие - легко интерпретируемые модели → инсайты

Основная задача – понять закономерности происходящих процессов, найдя зависимости в данных.

Акцент - на статистической значимости результатов.

Примеры: линейные модели, деревья решений, k-means кластеризация

Предиктивные – часто black box модели → высокая точность предсказаний

Предиктивные vs Объясняющие

Объясняющие - легко интерпретируемые модели → инсайты

Основная задача – понять закономерности происходящих процессов, найдя зависимости в данных.

Акцент - на статистической значимости результатов.

Примеры: линейные модели, деревья решений, k-means кластеризация

Предиктивные – часто black box модели → высокая точность предсказаний

Акцент – на точности и устойчивости предсказаний.

Предиктивные vs Объясняющие

Объясняющие - легко интерпретируемые модели → инсайты

Основная задача – понять закономерности происходящих процессов, найдя зависимости в данных.

Акцент - на статистической значимости результатов.

Примеры: линейные модели, деревья решений, k-means кластеризация

Предиктивные – часто black box модели → высокая точность предсказаний

Акцент – на точности и устойчивости предсказаний.

Примеры: градиентный бустинг, нейронные сети, многие рекомендашки.

#018

Train-test – зачем и как правильно?

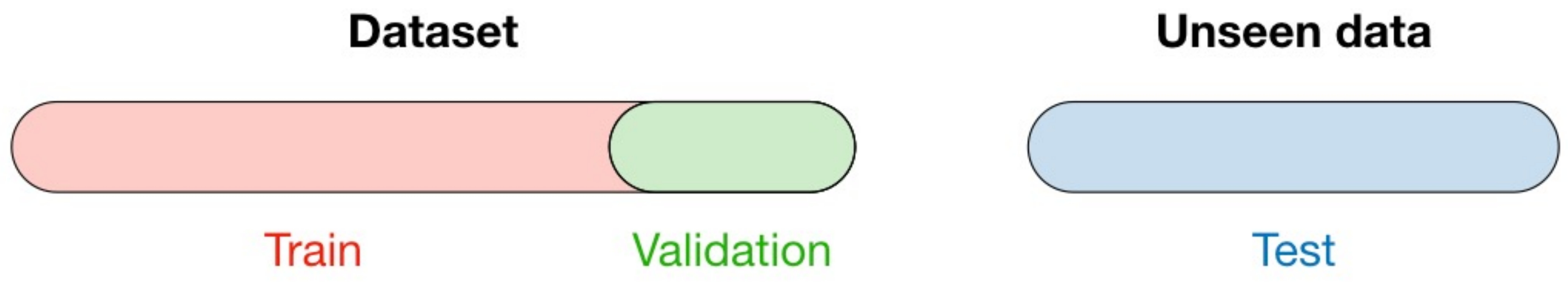
В процессе построения модели данные часто делят на две или т

- Train
- Validation
- [Test]

Train-test – зачем и как правильно?

В процессе построения модели данные часто делят на две или т

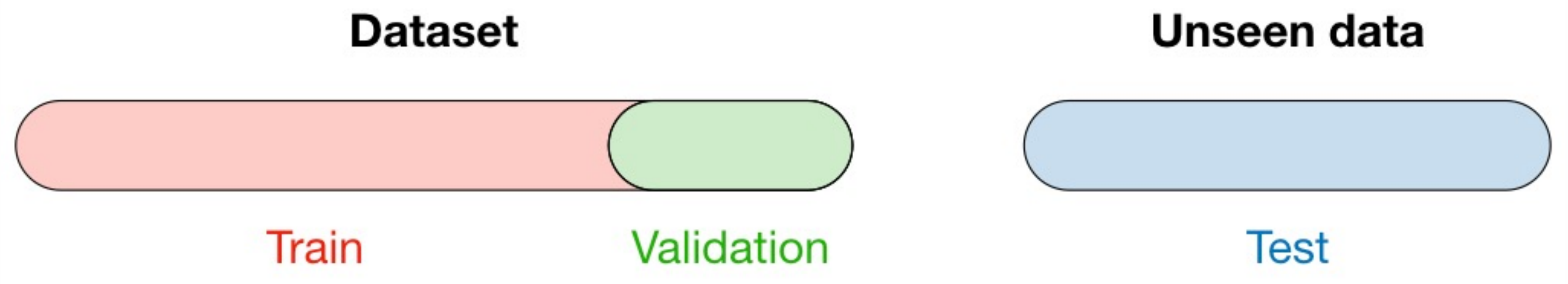
- Train
- Validation
- [Test]



Train-test – зачем и как правильно?

В процессе построения модели данные часто делят на две или т

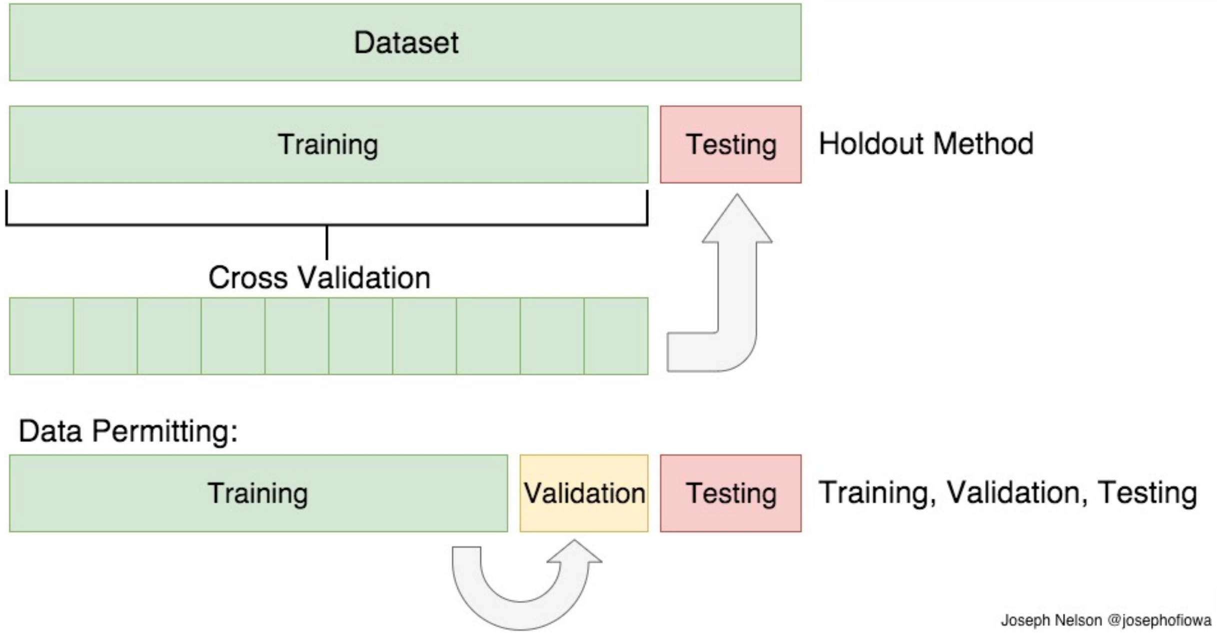
- Train
- Validation
- [Test]

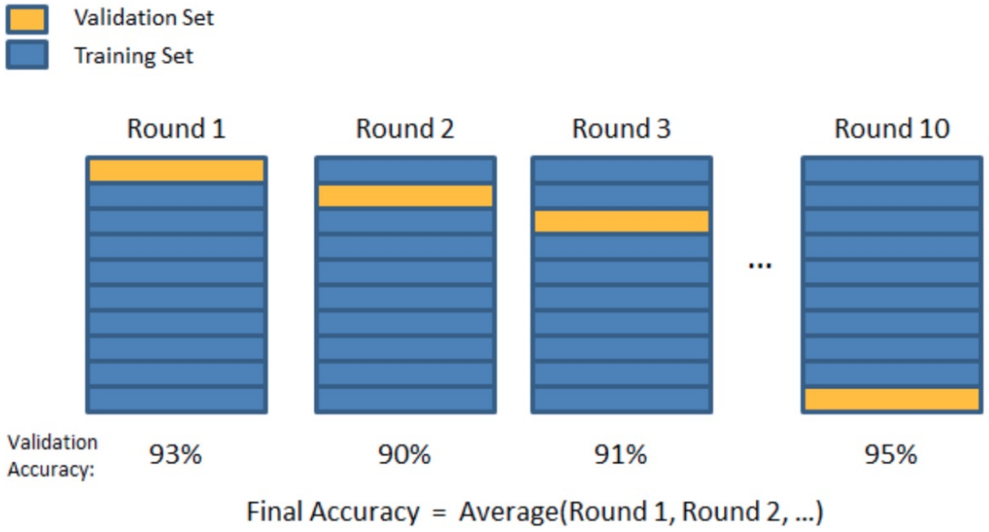


Зачем?

Train-test – зачем и как правильно?

Модификации и варианты:



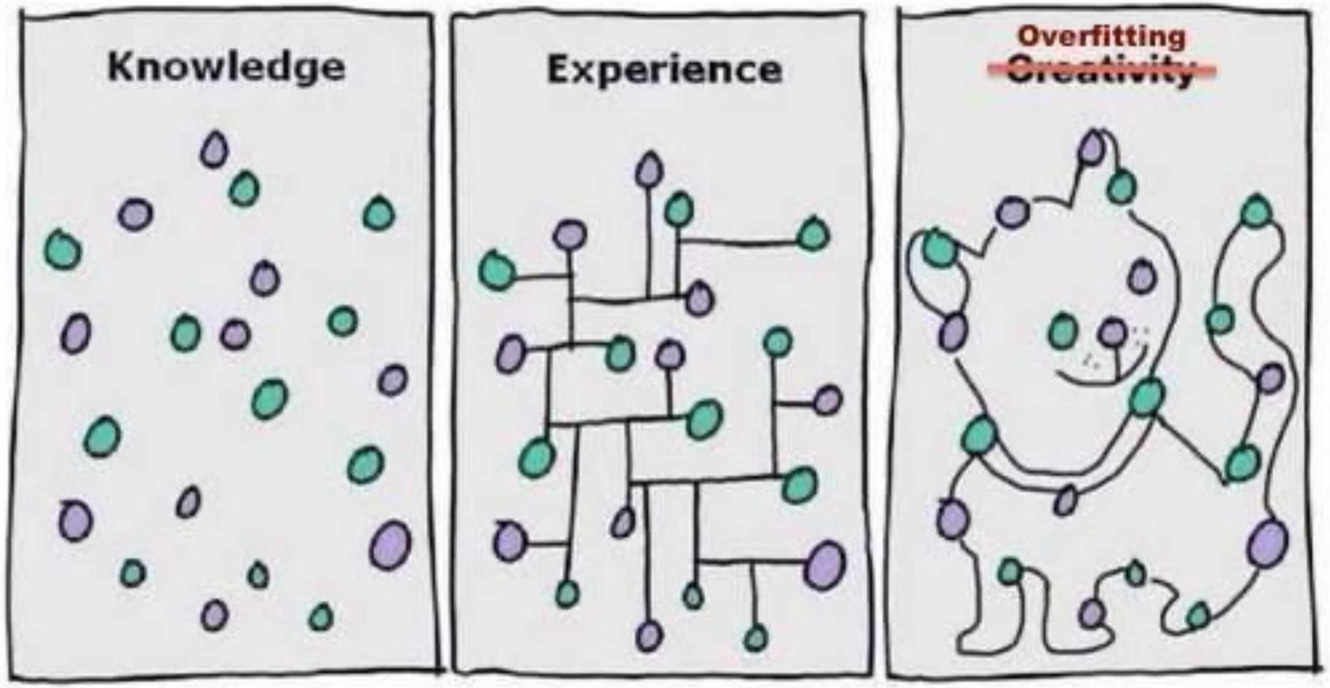


https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Так все-таки зачем?

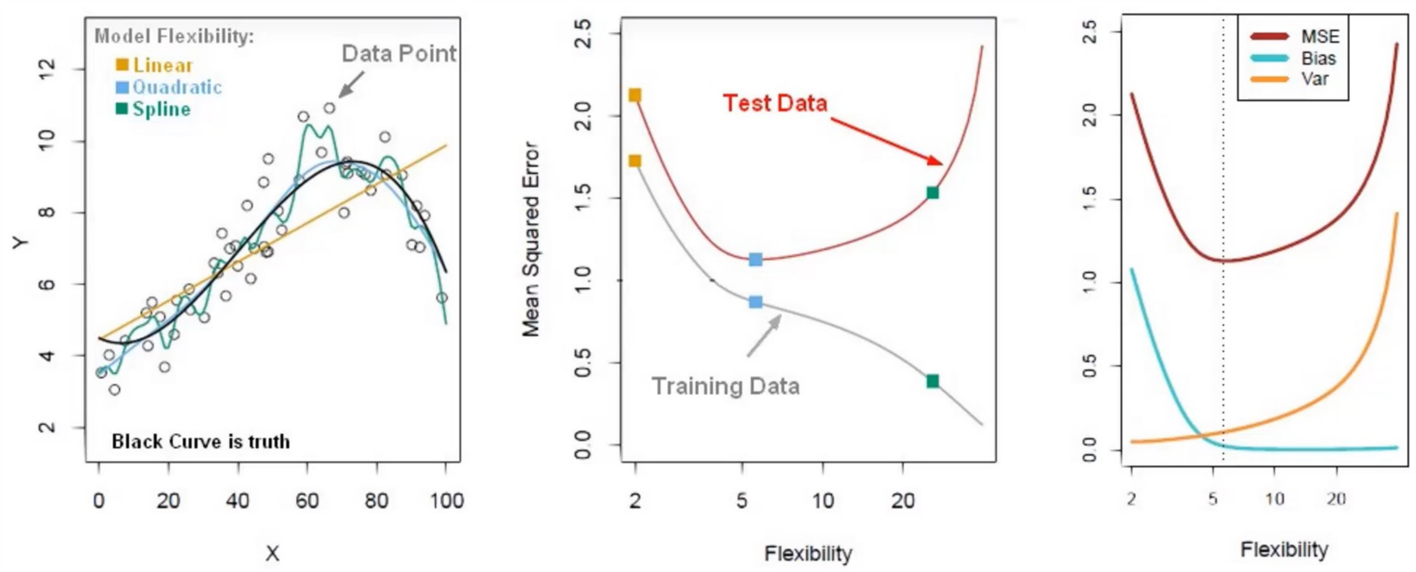
Так все-таки зачем?

Вот зачем:



Так все-таки зачем?

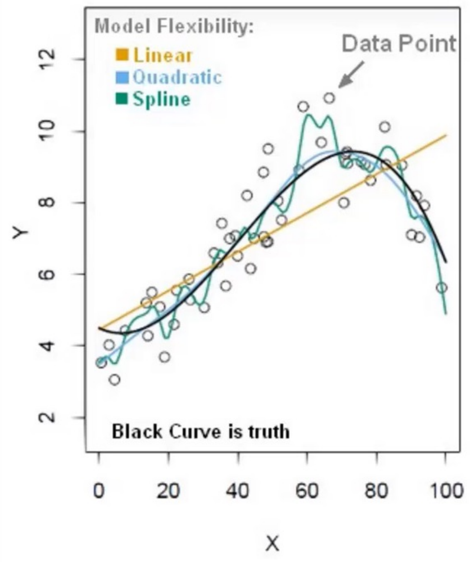
Другими словами:

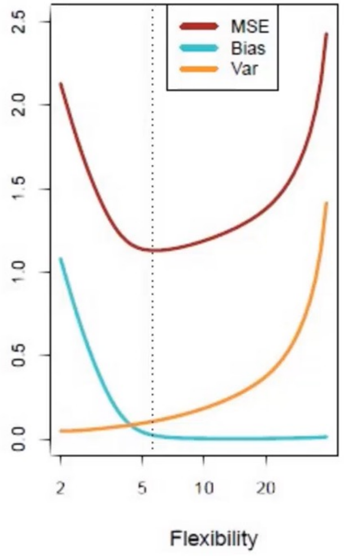


https://www.andreaperlato.com/theorypost/bias-variance-trade-off/

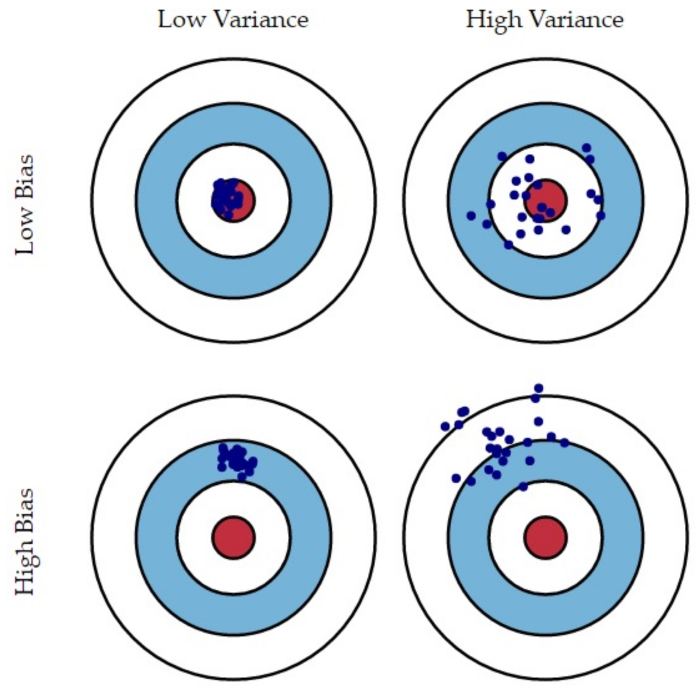
Так все-таки зачем?

Другими словами:





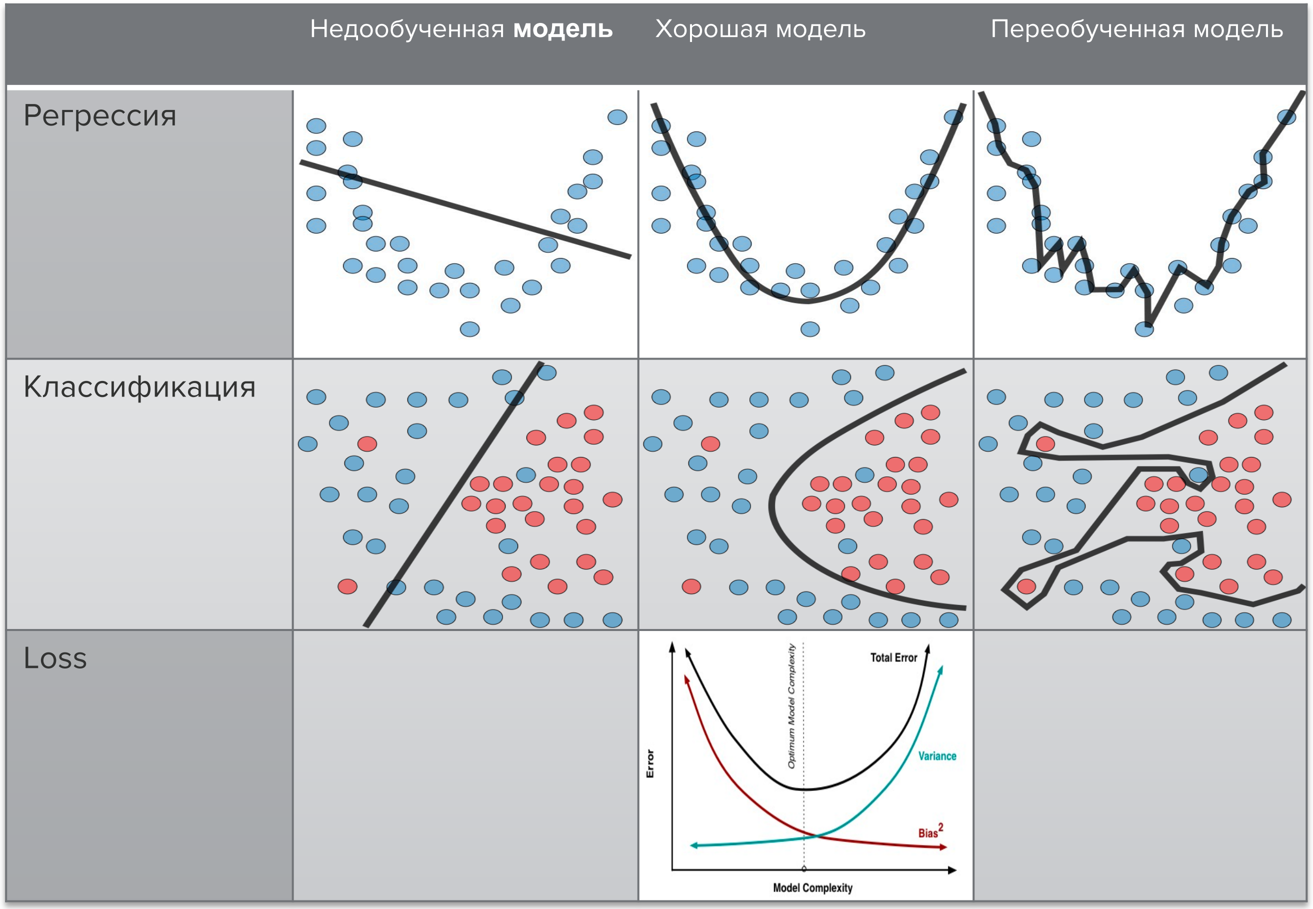
https://www.andreaperlato.com/theorypost/bias-variance-trade-off/



https://medium.com/@akgone38/what-the-heck-bias-variance-tradeoff-is-fe4681c0e71b

Так все-таки зачем?

Еще одна иллюстрация:



Так все-таки зачем?

Еще одна иллюстрация:

Моделечки!

Моделечки!

Моделечки!

Эльдар?

Моделечки!

Обучение с учителем

Обучение с учителем

- Задача регрессии
  Целевая переменная $y \in \{-\infty, \infty\}$

Обучение с учителем

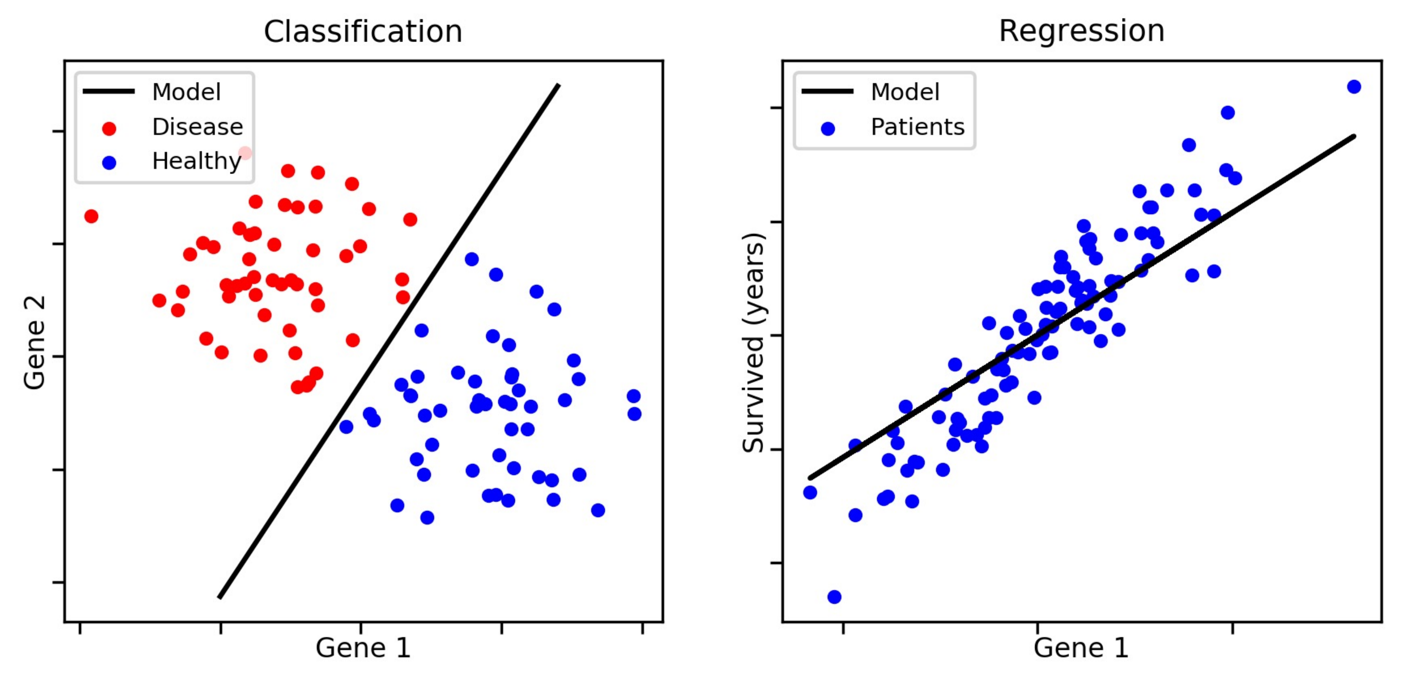
- Задача регрессии
  Целевая переменная $y \in \{-\infty, \infty\}$

- Задача классификации
  В общем случае $y \in \{1, 2, \ldots, K\}$, бинарная - $y \in \{0, 1\}$

Обучение с учителем

- Задача регрессии
  Целевая переменная $y \in \{-\infty, \infty\}$

- Задача классификации
  В общем случае $y \in \{1, 2, \dots, K\}$, бинарная - $y \in \{0, 1\}$



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 $Y \in \{1, \ldots, M\}$

Задача регрессии

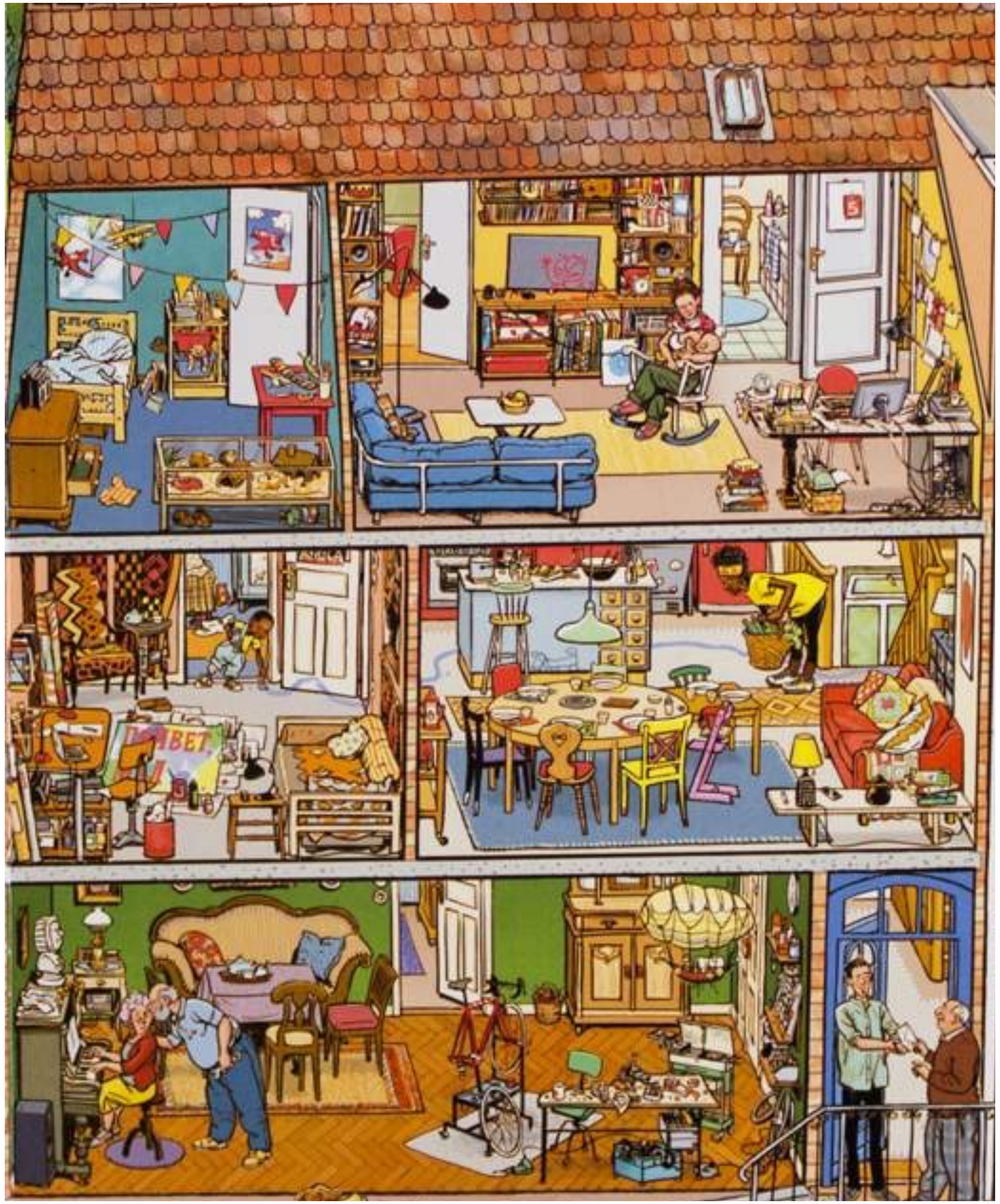
	Признаки (Features)							Target
	Площадь, м^2	Число комнат	Расстояние до центра, км	Новостройка	Наличие балкона	Время до метро, мин	Этаж	Высота потолков, м	Стоимость

Train

Площадь, м^2	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
…	…	…	…	…	…	…	…	…

Test

52	3	41	1	1	53	35	2.8	?



#039

Nearest Neighbors

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

Nearest Neighbors

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

- Для регрессии:

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

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

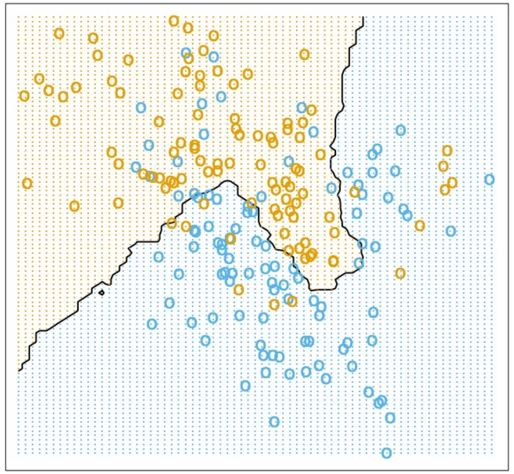
Непараметрический метод, требующий минимальное число предположений о характере функции f.

- 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

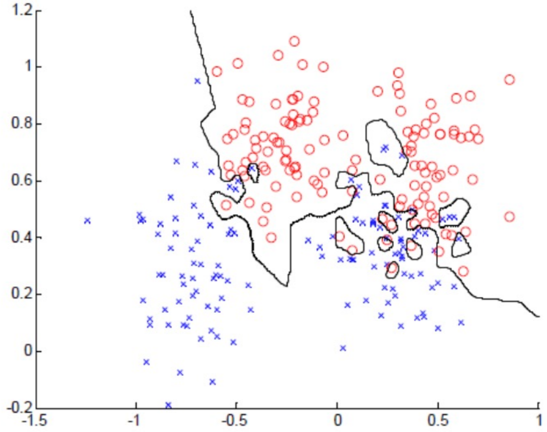
Nearest Neighbors

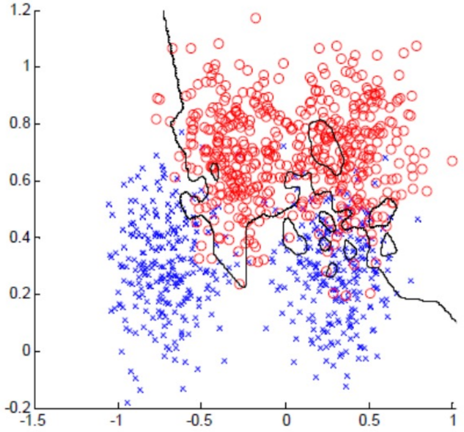
- Какие проблемы могут возникнуть?
- Выбор k
- Выбор метрики близости
- Скорость вычислений

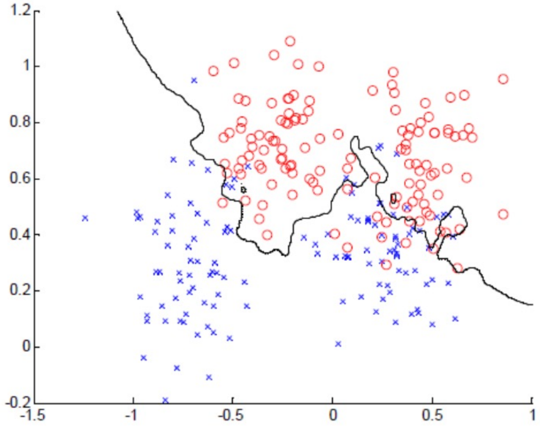
Nearest Neighbors

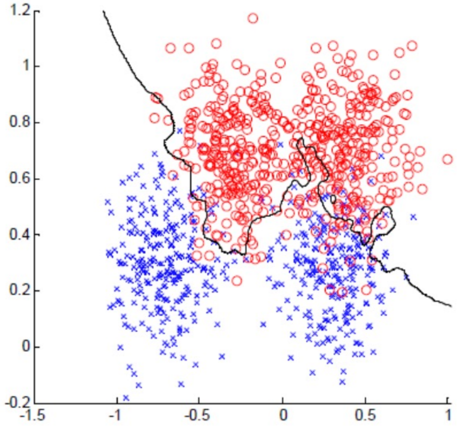
- Какие проблемы могут возникнуть?
- Выбор k
- Выбор метрики близости
- Скорость вычислений
- Проклятие размерности

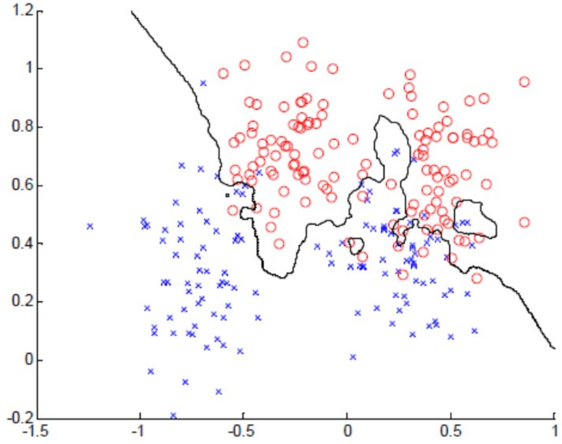
Nearest Neighbors

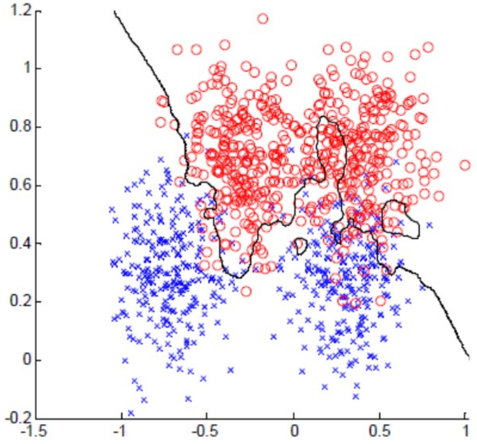












Какие значения k в данных примерах классификации?

Nearest Neighbors

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

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

Nearest Neighbors

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

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

Как называется данная метрика:

- При p=1?

Nearest Neighbors

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

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

Как называется данная метрика:

- При p=1?
- При p=2?

Nearest Neighbors

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

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

Как называется данная метрика:

- При p=1?
- При p=2?
- При $p \to \infty$? (Пафнутий)

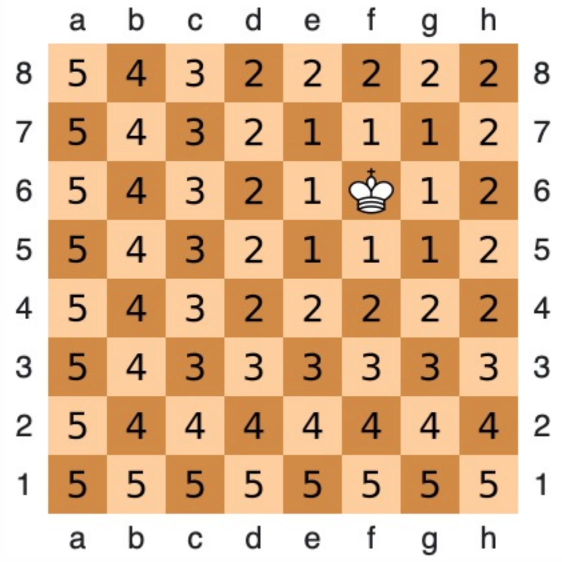
Nearest Neighbors

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

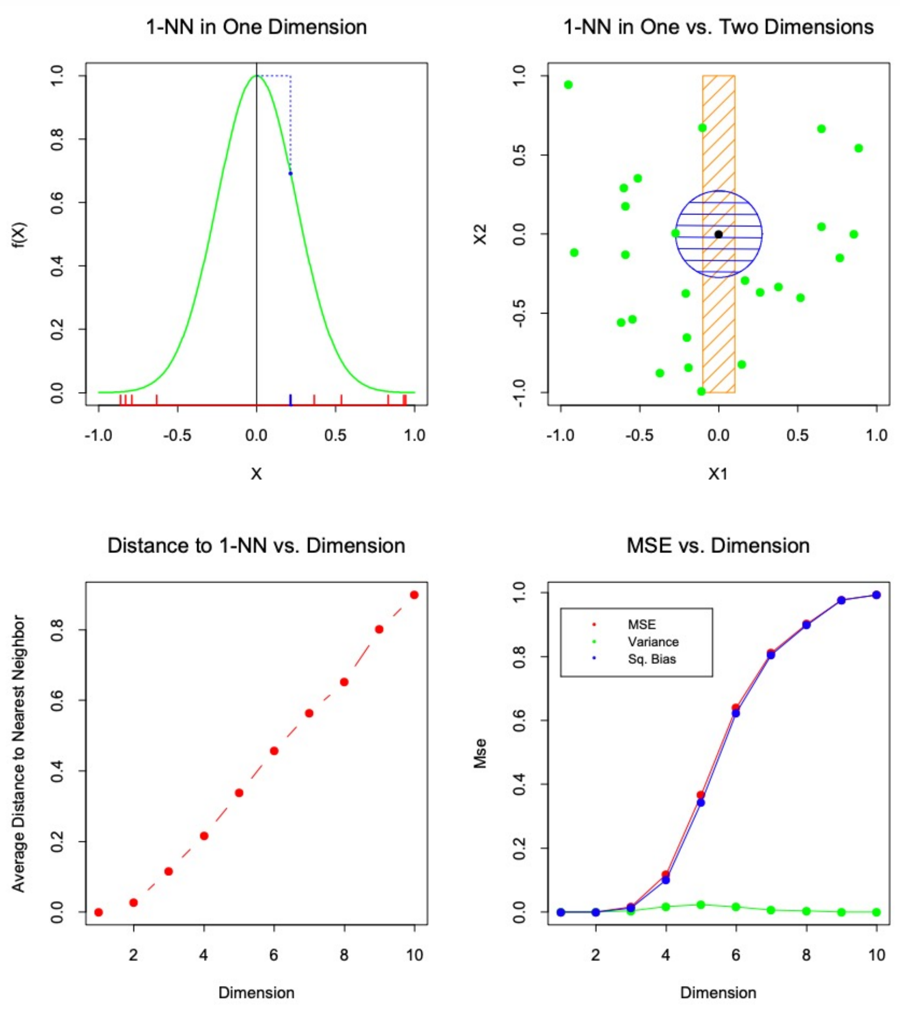
$$\mathrm{dist}(\mathbf{x}, \mathbf{z}) = \left(\sum_{r=1}^{d} |x_r - z_r|^p \right)^{1/p}$$

Как называется данная метрика:

- При p=1?
- При p=2?
- При $p \to \infty$? (Пафнутий)

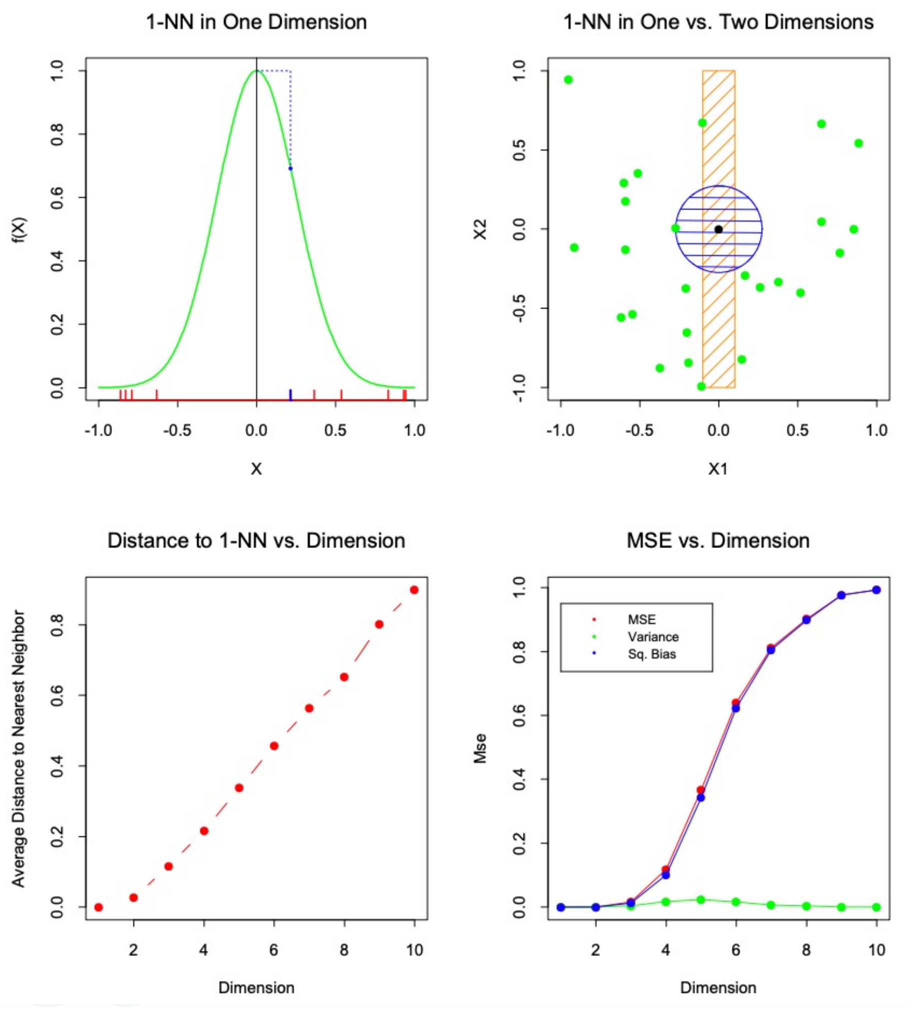


Nearest Neighbors



В чем состоит проклятие размерности для kNN?

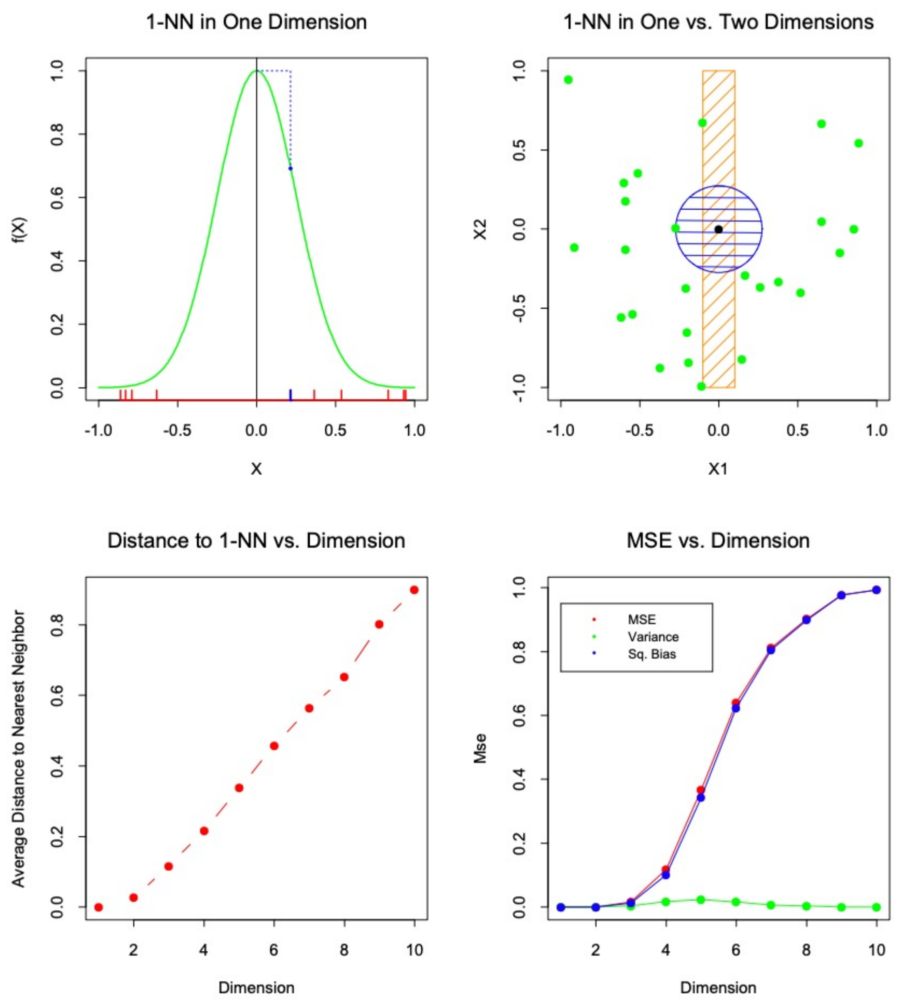
Nearest Neighbors



В чем состоит проклятие размерности для kNN?

- Расстояние до ближайшего соседа растет с ростом размерности.

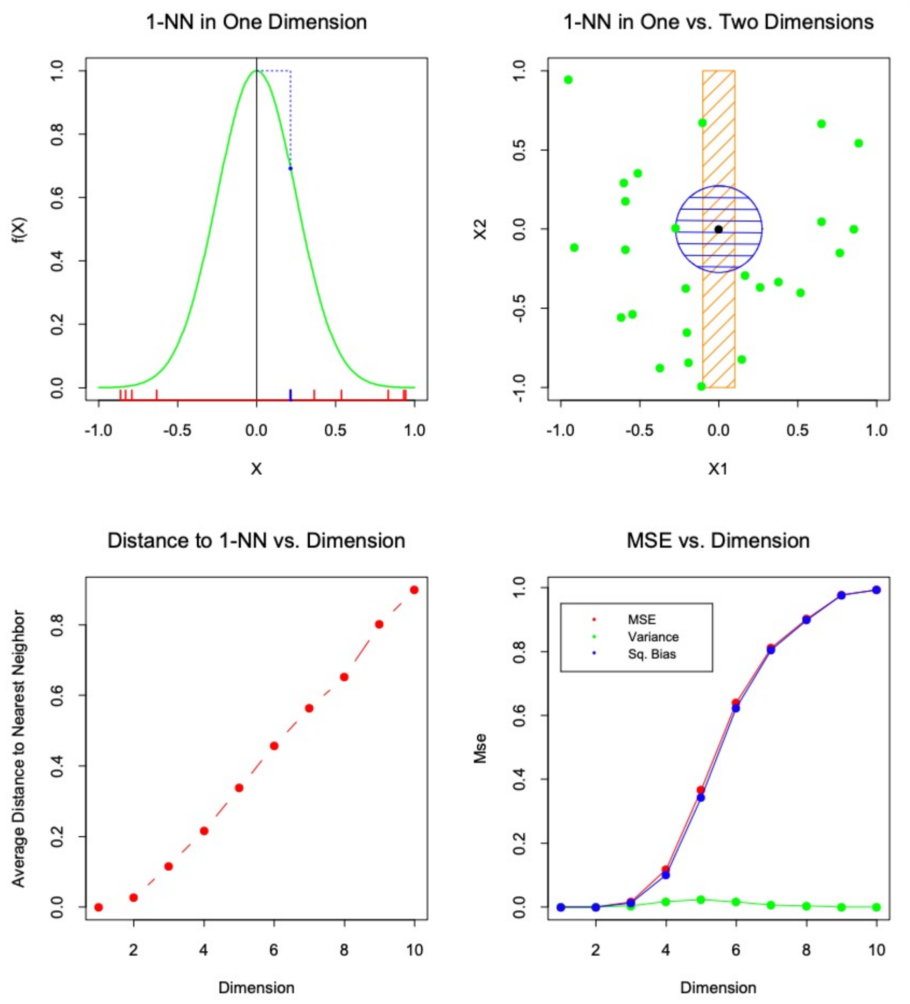
Nearest Neighbors



В чем состоит проклятие размерности для kNN?

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

Nearest Neighbors



В чем состоит проклятие размерности для kNN?

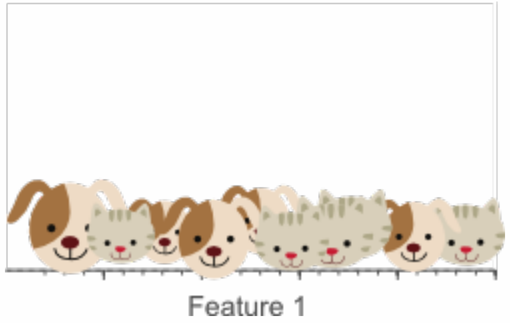
- Расстояние до ближайшего соседа растет с ростом размерности.
- По этой причине, растет bias нашей оценки.
- Более того, с ростом размерности основная масса данных при
определенных допущениях распределяется по поверхности гиперсферы,
что делает ближайших соседей практически неотличимыми от дальних.

Nearest Neighbors

Проклятие размерности $\lim\limits_{d\to\infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} \to 0$

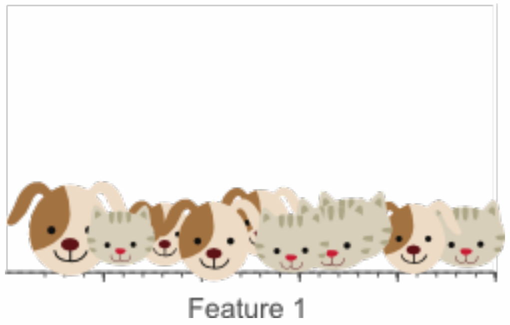
Nearest Neighbors

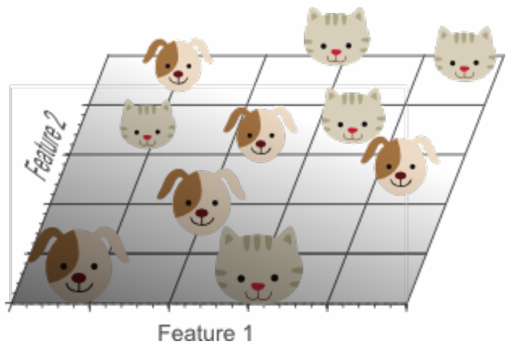
Проклятие размерности $\lim_{d \to \infty} \frac{\text{dist}_{\text{max}} - \text{dist}_{\text{min}}}{\text{dist}_{\text{min}}} \to 0$



Nearest Neighbors

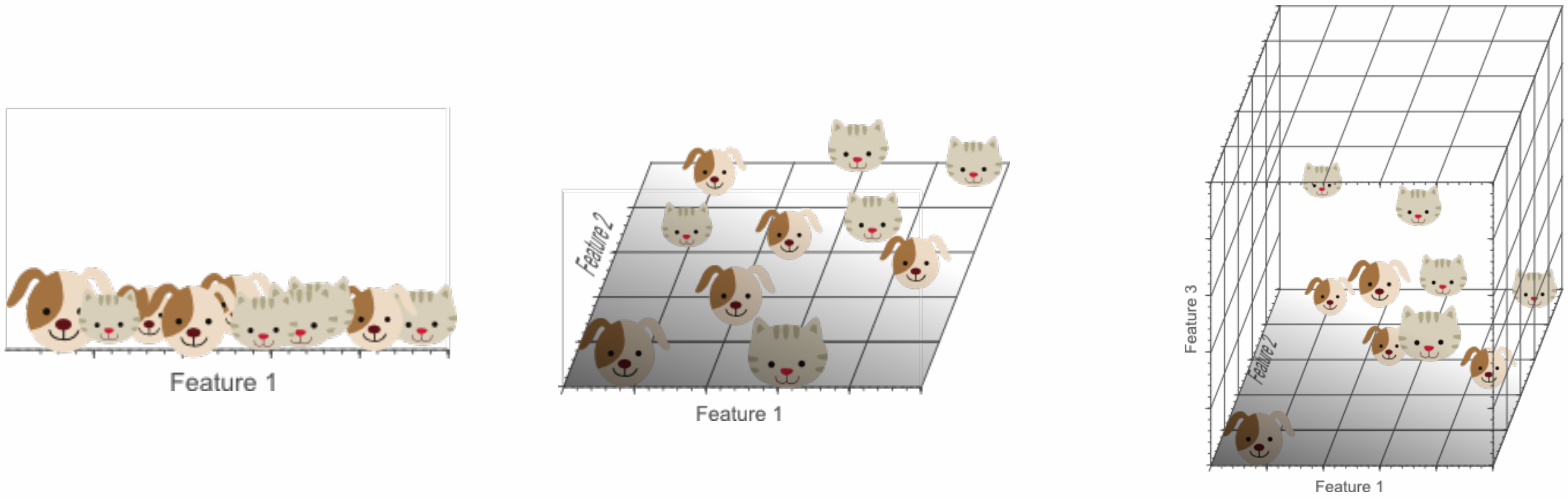
Проклятие размерности $\lim_{d \to \infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} \to 0$



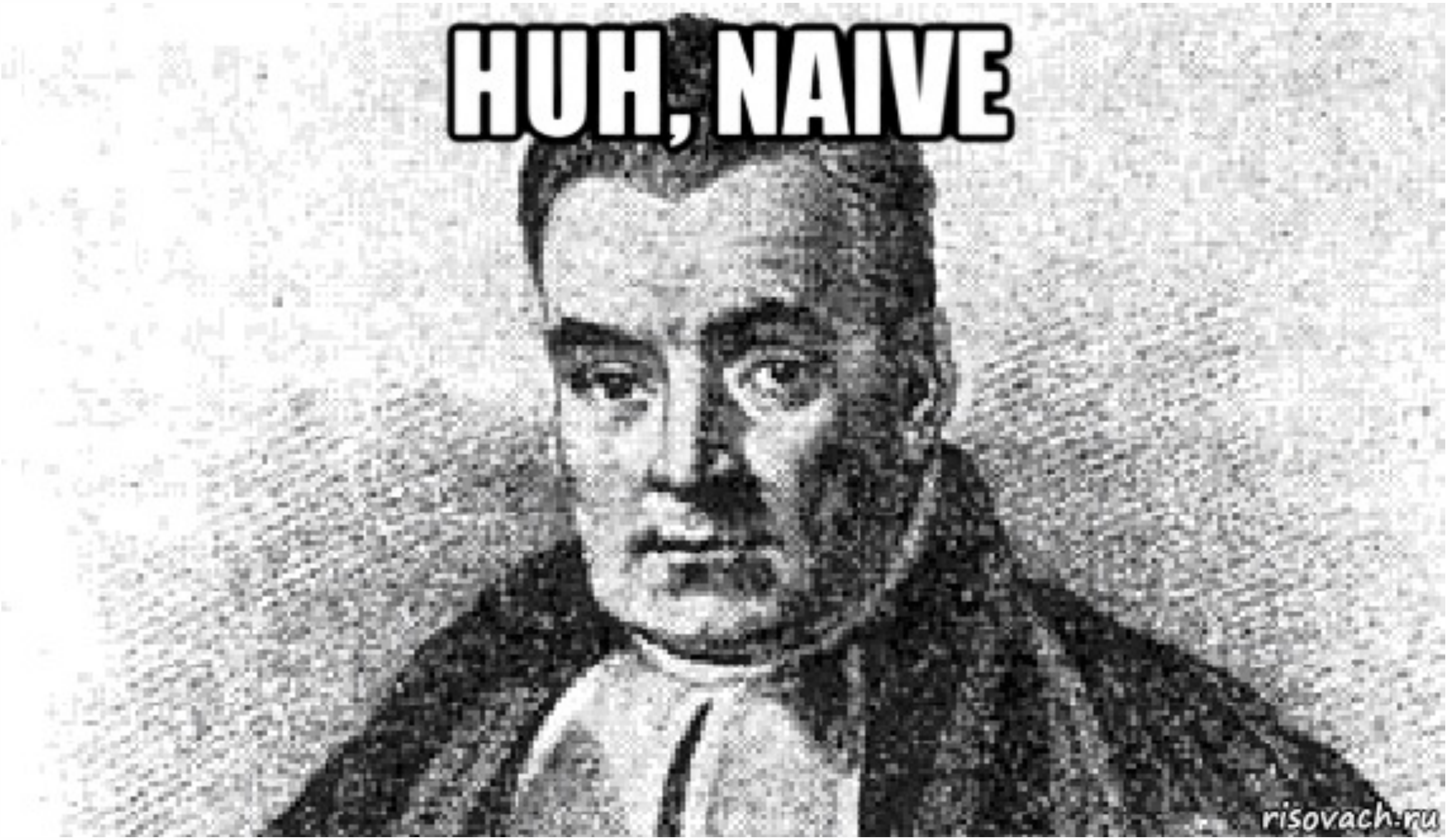


Nearest Neighbors

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



#062


HUH, NAIVE
risovach.ru

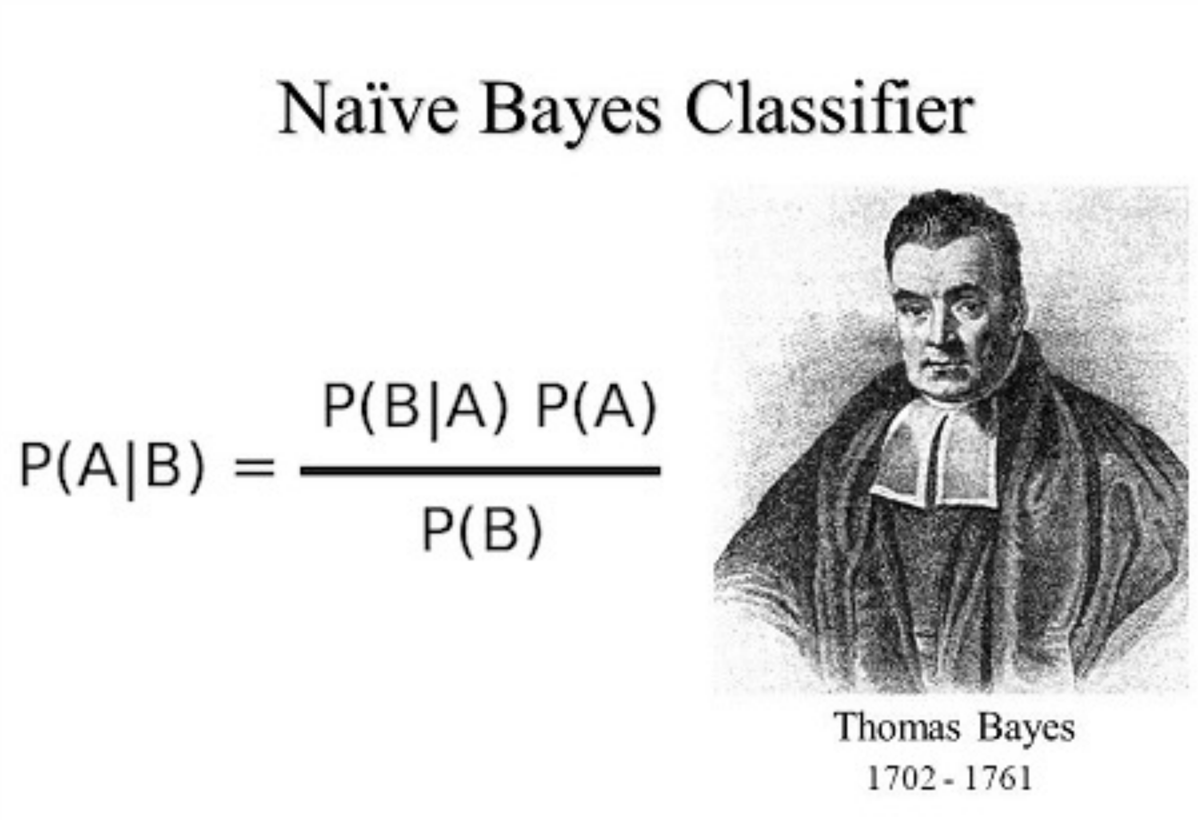
Naïve Bayes

Naïve Bayes Classifier

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

Thomas Bayes
1702 - 1761

Naïve Bayes



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 $X = \{x_1, x_2, \ldots, x_m\}$

Naïve Bayes

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

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

Naïve Bayes

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

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

Отсюда переходим к финальной проблеме:

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

$$= arg\max_y P(X|y)P(y) =$$

$$= arg\max_y P(y)\prod_i P(x_i|y)$$

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!

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

Линейная регрессия - линейная модель регрессии, которая предполагает определенную связь между зависимой и независимыми переменными, а именно:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

Здесь ϵ_i - случайная величина, независимо одинаково распределенная для каждого наблюдения и
отклоняющая значение игрек от линейной модели.

Linear models. Постановка задачи

Линейная регрессия - линейная модель регрессии, которая предполагает определенную связь между зависимой и независимыми переменными, а именно:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

Здесь ϵ_i - случайная величина, независимо одинаково распределенная для каждого наблюдения и
отклоняющая значение игрек от линейной модели.

Linear models. Постановка задачи

Случай двух переменных:

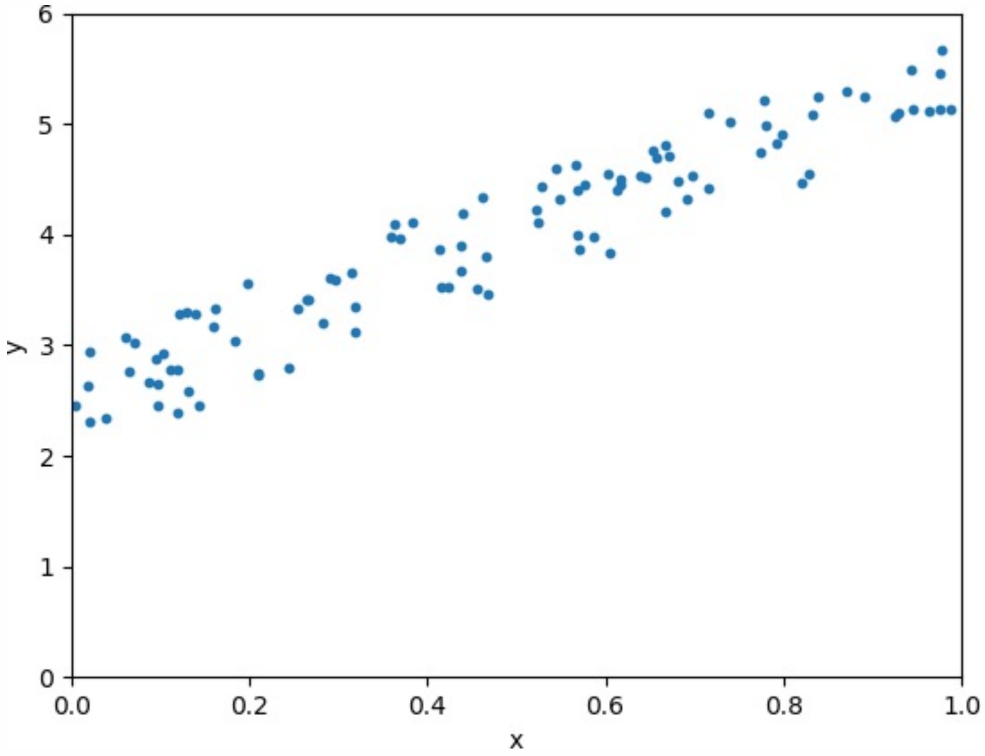


Рис.: Данные

Linear models. Постановка задачи

Случай двух переменных:

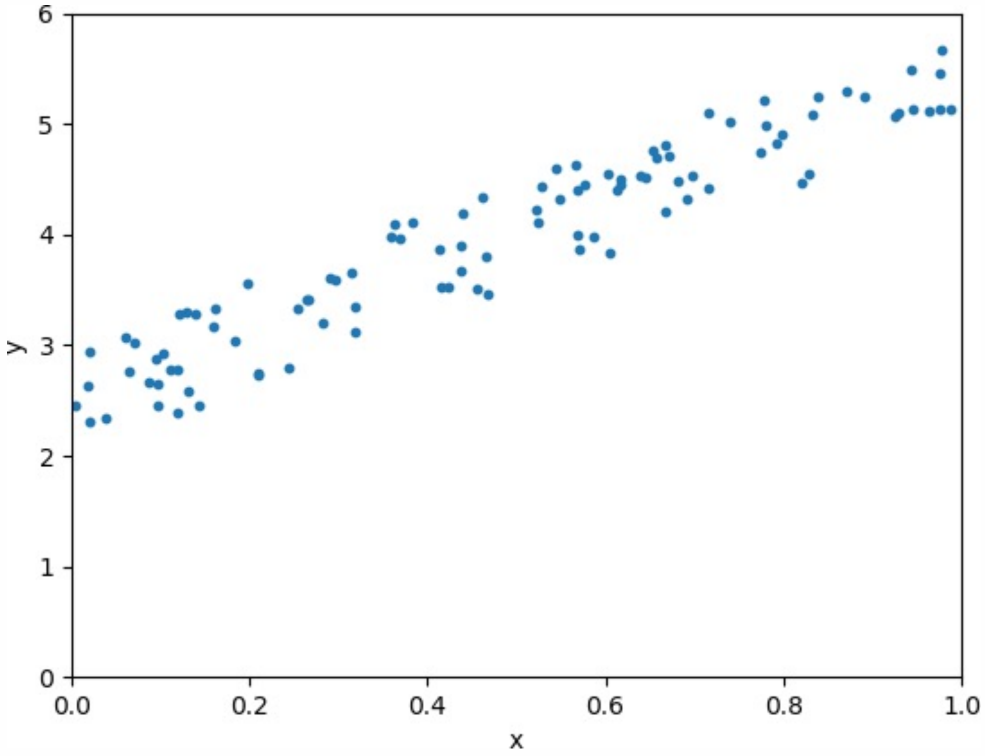


Рис.: Данные

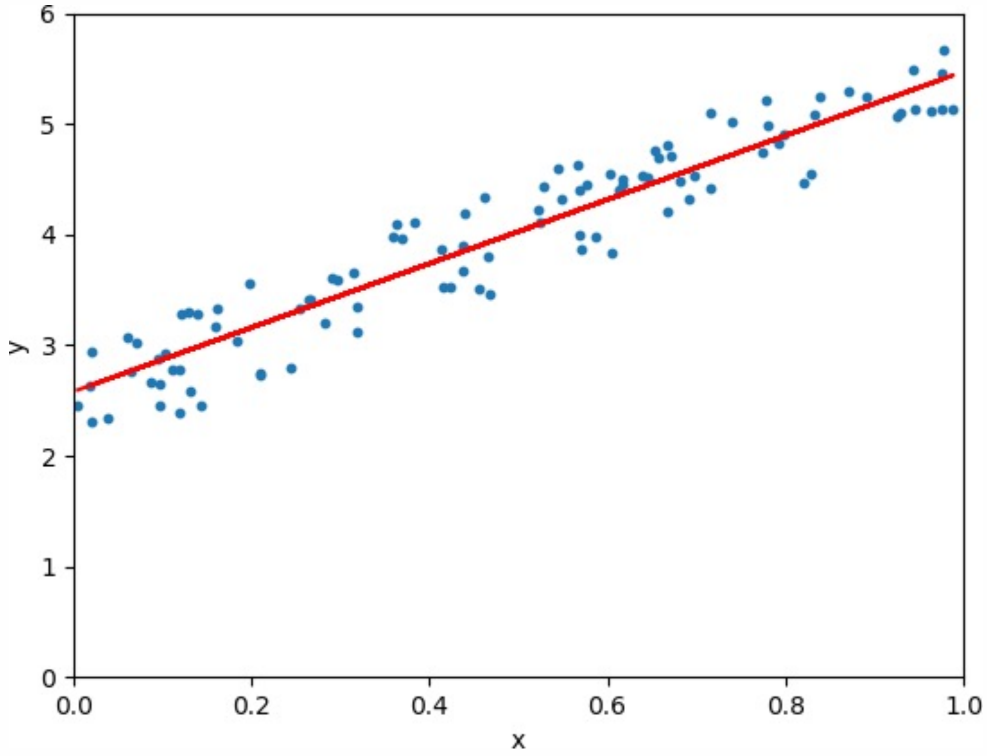


Рис.: Данные и линейная модель

Linear models. Постановка задачи

Случай двух переменных:

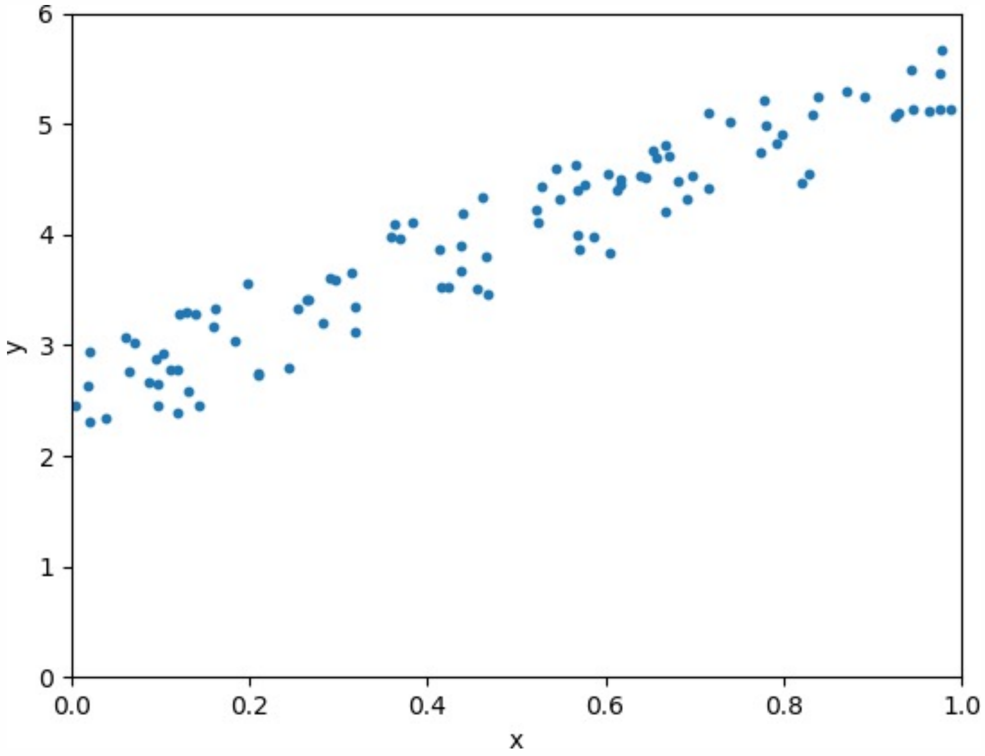


Рис.: Данные

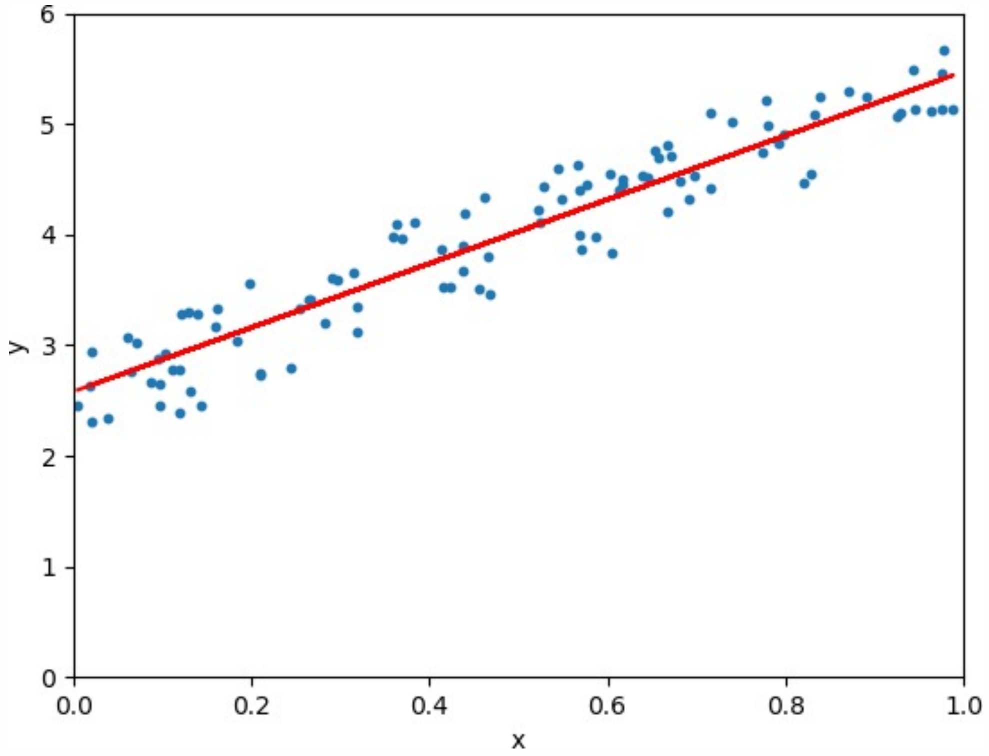


Рис.: Данные и линейная модель

Линейная модель линейна по параметрам!

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

Linear models. Постановка задачи

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

$$y = \hat{\beta}_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

Linear models. Постановка задачи

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

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

- Наша задача - получить оценку параметров модели β_i
  В итоге мы хотим получить уравнение:

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

Главная задача, которую мы здесь решаем – восстановить зависимость таргета y от известных нам признаков 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:

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

Linear models. CEF.

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

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

Для красоты:

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

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:

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

Linear models. OLS.

- Мы предположили, что данные генерируются линейной моделью:

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

Linear models. OLS.

- Мы предположили, что данные генерируются линейной моделью:

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

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

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

Linear models. Еще немного.

Здесь у нас MSE в качестве функции потерь:

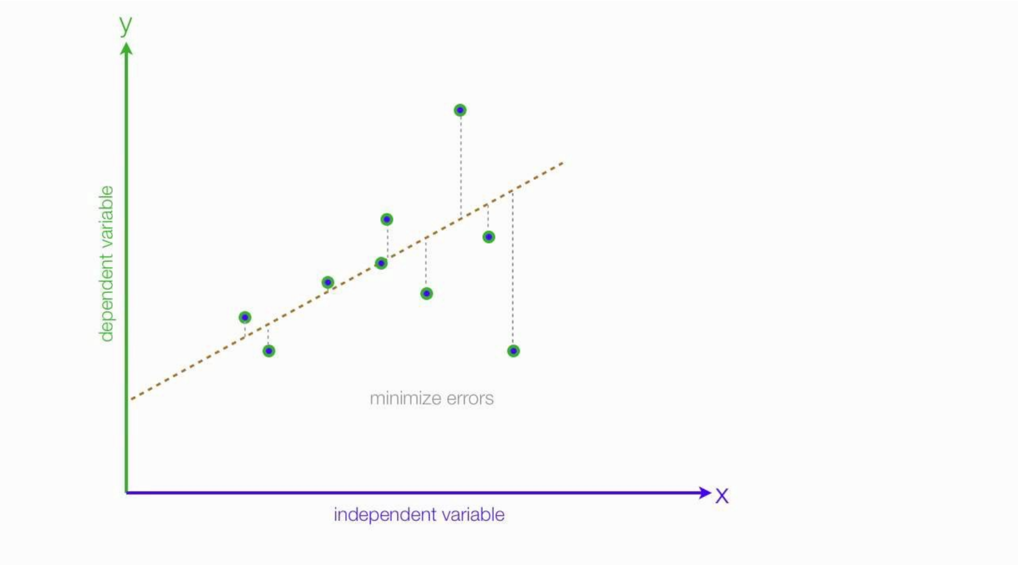
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Linear models. Еще немного.

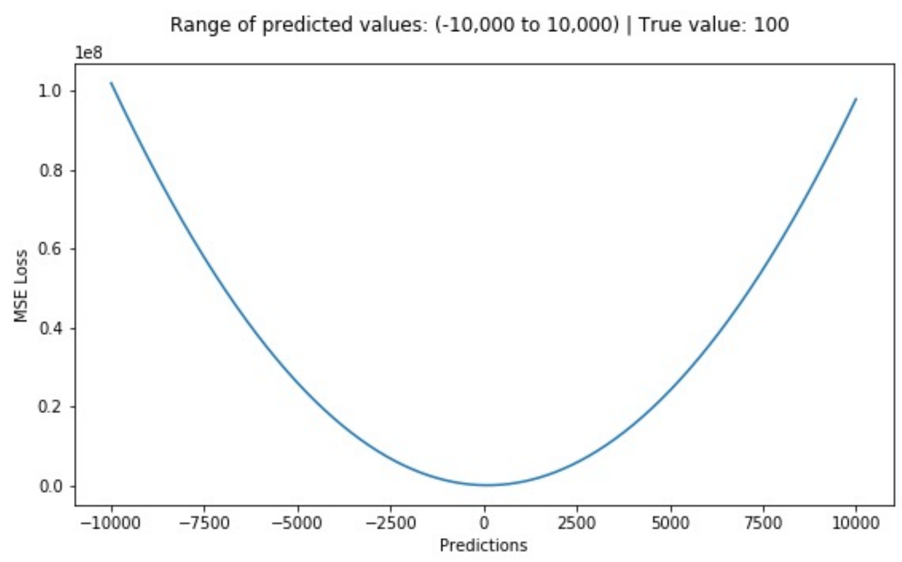
Здесь у нас MSE в качестве функции потерь:

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

Тогда задача и функция потерь выглядят:



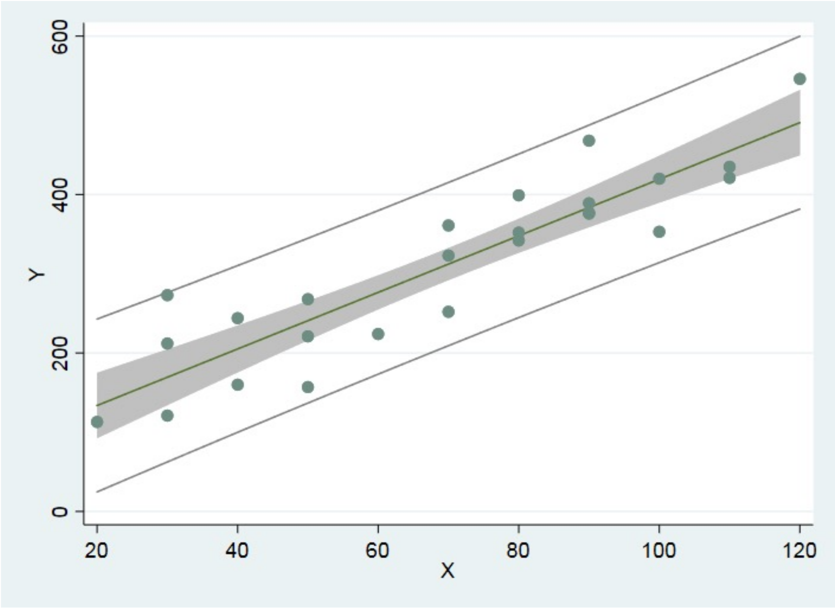
https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e



https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0

Linear models. Еще немного.

А решение будет что-то вроде этого:



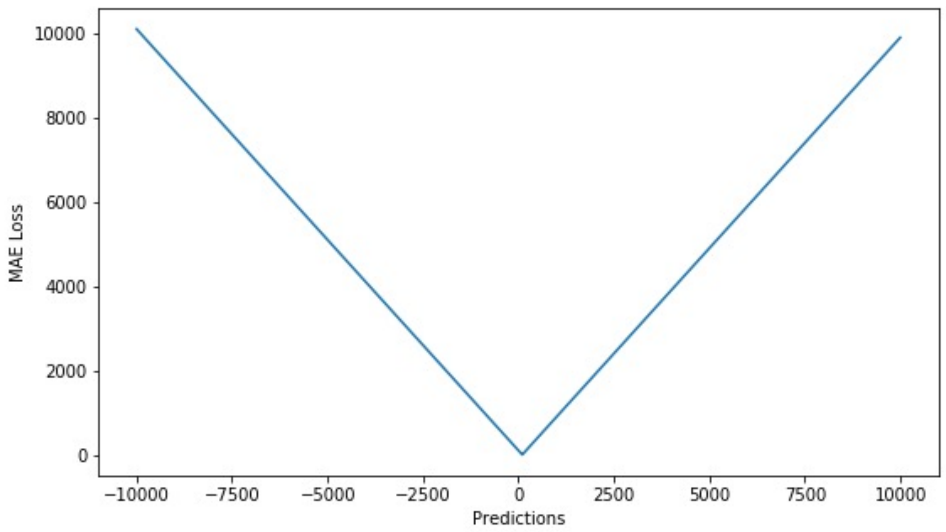
Linear models. Еще немного.

Существуют и другие функции потерь. Например:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Задача будет та же, только потери оптимизируем теперь другие:

Range of predicted values: (-10,000 to 10,000) | True value: 100



#094

Linear models. Bonus slide. Assumptions OLS.

Assumtion 1 (экзогенность):

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

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

Linear models. Bonus slide. Assumptions OLS.

Assumtion 1 (экзогенность):

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

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

Assumtion 2 (полный ранг):

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

Эквивалентно предположению о положительно определенной матрице.

Linear models. Bonus slide. Assumptions OLS.

Assumtion 1 (экзогенность):

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

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

Assumtion 2 (полный ранг):

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

Эквивалентно предположению о положительно определенной матрице.

Теорема (состоятельность МНК):

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

Linear models. Bonus slide. Assumptions OLS.

Assumtion 3 (гомоскедастичность):

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

Где $\sigma^2 = E(\epsilon^{\mathbf{2}})$.

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. Bonus slide. Assumptions OLS.

Assumtion 3 (гомоскедастичность):

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

Где $\sigma^2 = E(\epsilon^{\mathbf{2}})$.

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:

$$f(y_1, \ldots, y_n | \mathbf{x_1}, \ldots, \mathbf{x_n}) = \prod_{i=1}^{n} f(y_i | \mathbf{x_i})$$

Linear models. ММП.

Записывая условную плотность для всей выборки, получим функцию правдоподобия:

$$f(y_1, ..., y_n | \mathbf{x_1}, ..., \mathbf{x_n}) = \prod_{i=1}^{n} f(y_i | \mathbf{x_i})$$

Логарифмируя и находя максимум из FOC:

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

Linear models. Проверка гипотез.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Calorie_Burnage   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.006
Method:                 Least Squares   F-statistic:                   0.04975
Date:                Tue, 29 Sep 2020   Prob (F-statistic):              0.824
Time:                        17:48:00   Log-Likelihood:                -1145.8
No. Observations:                 163   AIC:                             2296.
Df Residuals:                     161   BIC:                             2302.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept       346.8662    160.615      2.160      0.032      29.682     664.050
Average_Pulse     0.3296      1.478      0.223      0.824      -2.588       3.247
==============================================================================
Omnibus:                      124.542   Durbin-Watson:                   1.620
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              938.541
Skew:                           2.927   Prob(JB):                    1.58e-204
Kurtosis:                      13.195   Cond. No.                         811.
==============================================================================
```

https://www.w3schools.com/datascience/ds_linear_regression_coeff.asp

Linear models. Проверка гипотез.

Оценка коэффициента регрессии – случайная величина.
При выполнении предположений она имеет (асимптотически) нормальное распределение с математическим ожиданием равным реальному коэффициенту регрессии и некоторой дисперсией.

Смотрим на t-статистику и её p-value.

std err	t	P>|t|	[0.025	0.975]
160.615	2.160	0.032	29.682	664.050
1.478	0.223	0.824	-2.588	3.247

https://www.w3schools.com/datascience/ds_linear_regression_pvalue.asp

Linear models. Проверка гипотез.

t – статистика здесь позволяет проверить гипотезу:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

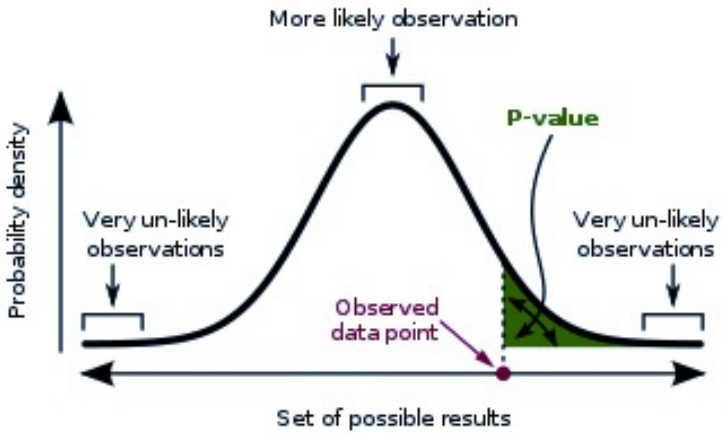
Linear models. Проверка гипотез.

t – статистика здесь позволяет проверить гипотезу:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

Распределение t-статистики при справедливости нулевой гипотезы:

$$t = \frac{\hat{\beta}}{se(\hat{\beta})} \sim t(n-k-1)|_{H_0}$$



https://data-flair.training/blogs/python-statistics/

Linear models. Проверка гипотез.

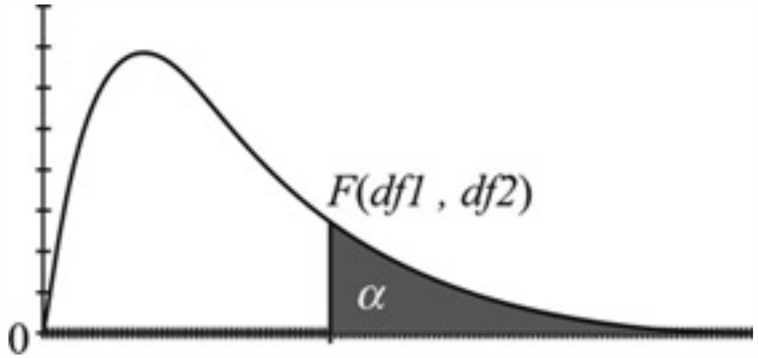
F – статистика:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : not\ H_0$$

Сама F статистика выглядит следующим образом:

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\left(\frac{RSS_2}{n - p_2}\right)} \sim F(p_2 - p_1, n - p_2)$$



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:

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

.

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

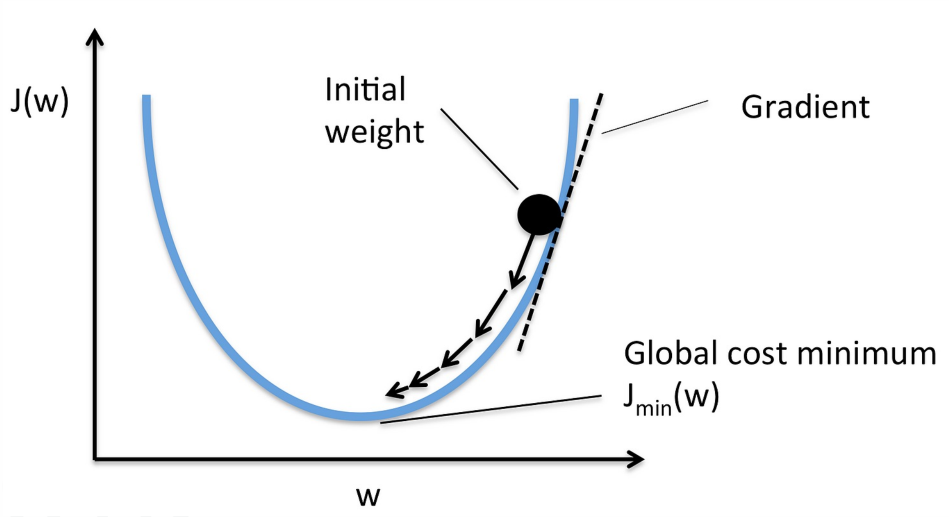
$$|\hat{\beta}^{(i+1)} - \hat{\beta}^{(i)}| < \epsilon$$

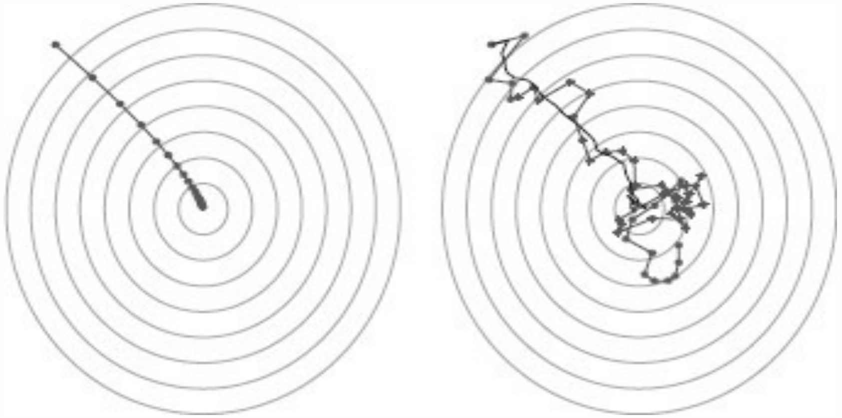
Linear models. Стохастический градиентный спуск.

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. 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 $k >> n$
- 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 $k >> n$
- 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?

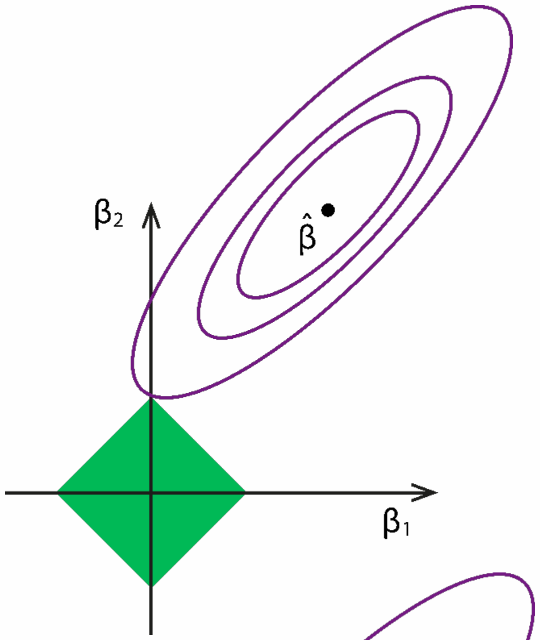
- Overfitting
- Feature selection
- High variance

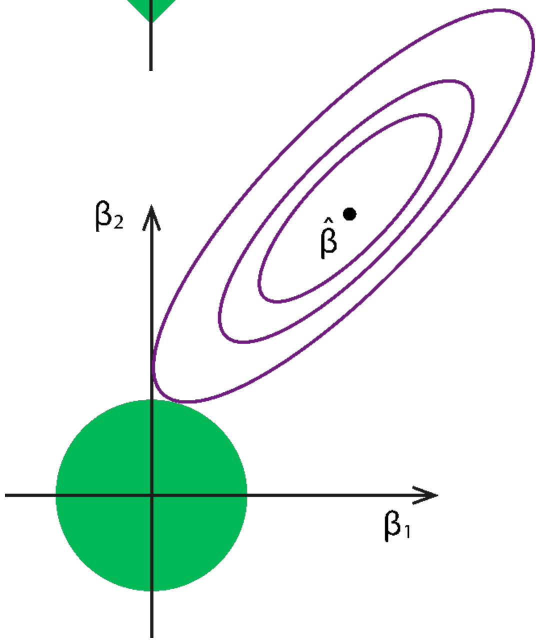
Linear models. Regularization.

Какие проблемы может решать регуляризация?

- Overfitting
- Feature selection
- High variance

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





Linear models. Classification.

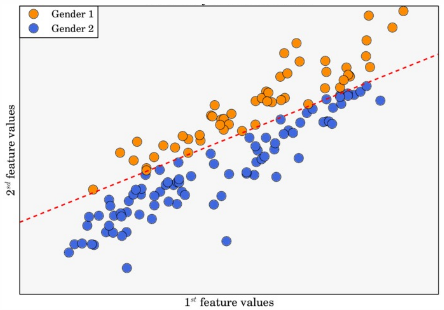
Задача, которую мы решаем в проблеме классификации может также быть разбита на два типа:

- Объяснить почему тот или иной объект принадлежит именно этому классу, а не другому
- Спрогнозировать – какому классу будет принадлежать объект с определенными характеристиками

Linear models. Classification.

Задача, которую мы решаем в проблеме классификации может также быть разбита на два типа:

- Объяснить почему тот или иной объект принадлежит именно этому классу, а не другому
- Спрогнозировать – какому классу будет принадлежать объект с определенными характеристиками

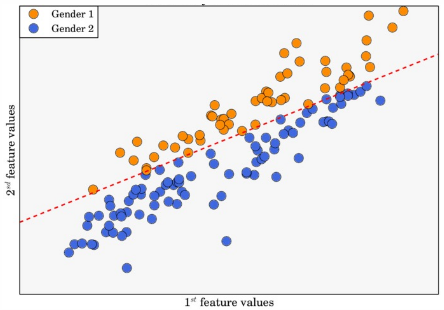


https://www.researchgate.net/figure/This-illustration-present-a-binary-classification-that-is-performed-on-two-features_fig5_285653348

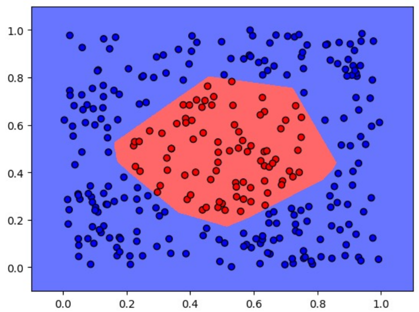
Linear models. Classification.

Задача, которую мы решаем в проблеме классификации может также быть разбита на два типа:

- Объяснить почему тот или иной объект принадлежит именно этому классу, а не другому
- Спрогнозировать – какому классу будет принадлежать объект с определенными характеристиками



https://www.researchgate.net/figure/This-illustration-present-a-binary-classification-that-is-performed-on-two-features_fig5_285653348

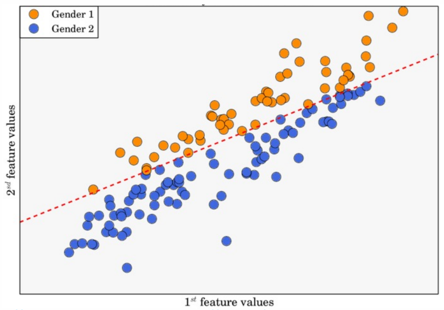


https://cfml.se/blog/binary_classification/

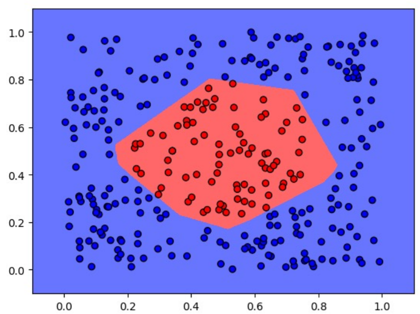
Linear models. Classification.

Задача, которую мы решаем в проблеме классификации может также быть разбита на два типа:

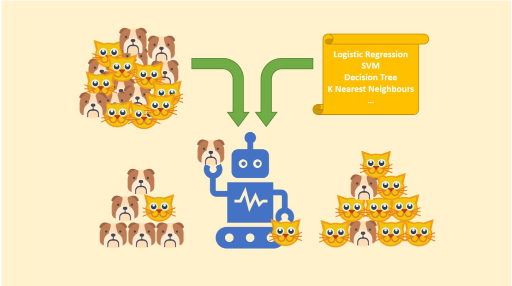
- Объяснить почему тот или иной объект принадлежит именно этому классу, а не другому
- Спрогнозировать – какому классу будет принадлежать объект с определенными характеристиками



https://www.researchgate.net/figure/This-illustration-present-a-binary-classification-that-is-performed-on-two-features_fig5_285653348



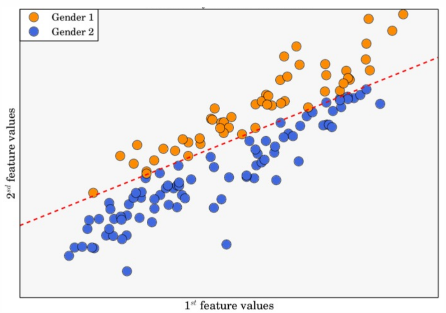
https://cfml.se/blog/binary_classification/



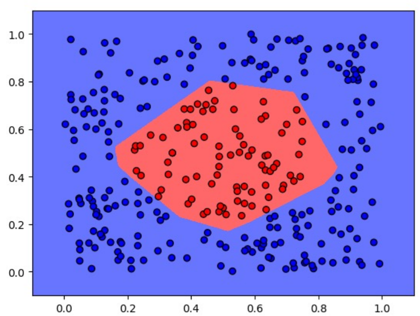
https://towardsdatascience.com/analytics-building-blocks-binary-classification-d205890314fc

Linear models. Classification.

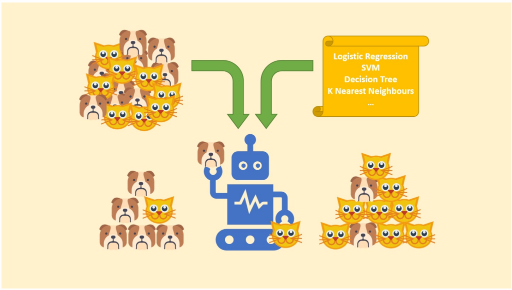
Задача, которую мы решаем в проблеме классификации может также быть разбита на два типа:
- Объяснить почему тот или иной объект принадлежит именно этому классу, а не другому
- Спрогнозировать – какому классу будет принадлежать объект с определенными характеристиками



https://www.researchgate.net/figure/This-illustration-present-a-binary-classification-that-is-performed-on-two-features_fig5_285653348



https://cfml.se/blog/binary_classification/



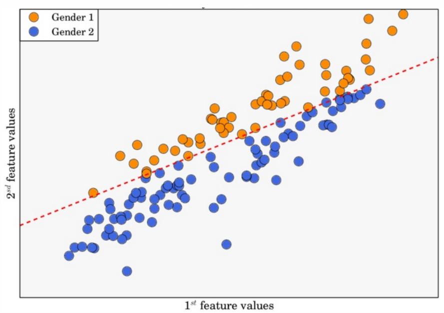
https://towardsdatascience.com/analytics-building-blocks-binary-classification-d205890314fc

Классификация может быть бинарная или мультиклассовая:
Примеры:
- Предсказание дефолта заемщика
- Предсказание направления движения акции
- Таргетирование маркетинговых коммуникаций
- К кому из 4х операторов подключится человек

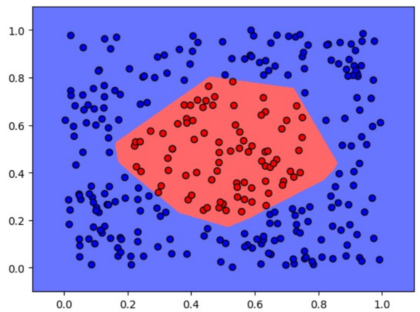
Linear models. Classification.

Задача, которую мы решаем в проблеме классификации может также быть разбита на два типа:

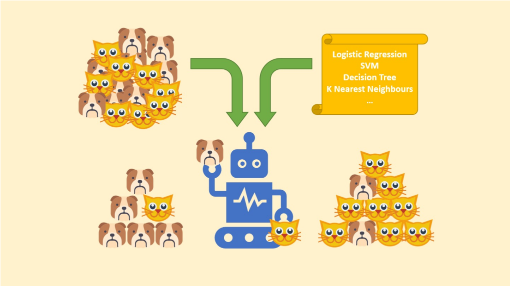
- Объяснить почему тот или иной объект принадлежит именно этому классу, а не другому
- Спрогнозировать – какому классу будет принадлежать объект с определенными характеристиками



https://www.researchgate.net/figure/This-illustration-present-a-binary-classification-that-is-performed-on-two-features_fig5_285653348



https://cfml.se/blog/binary_classification/



https://towardsdatascience.com/analytics-building-blocks-binary-classification-d205890314fc

Классификация может быть бинарная или мультиклассовая:

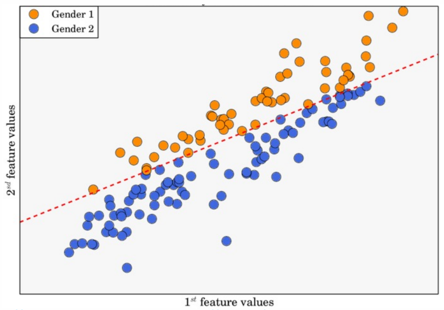
Примеры:

- Предсказание дефолта заемщика
- Предсказание направления движения акции
- Таргетирование маркетинговых коммуникаций
- К кому из 4х операторов подключится человек

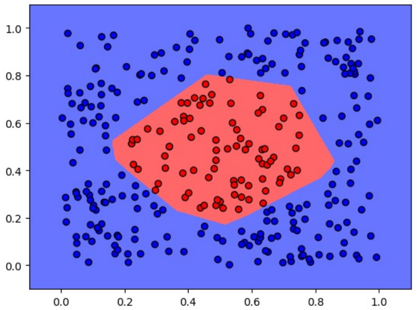
Linear models. Classification.

Задача, которую мы решаем в проблеме классификации может также быть разбита на два типа:

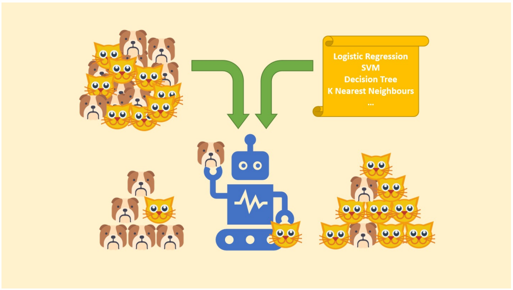
- Объяснить почему тот или иной объект принадлежит именно этому классу, а не другому
- Спрогнозировать – какому классу будет принадлежать объект с определенными характеристиками



https://www.researchgate.net/figure/This-illustration-present-a-binary-classification-that-is-performed-on-two-features_fig5_285653348



https://cfml.se/blog/binary_classification/



https://towardsdatascience.com/analytics-building-blocks-binary-classification-d205890314fc

Классификация может быть бинарная или мультиклассовая:

Примеры:

- Предсказание дефолта заемщика
- Предсказание направления движения акции
- Таргетирование маркетинговых коммуникаций
- К кому из 4х операторов подключится человек

Linear models. Classification.

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

Как будем решать?

Linear models. Classification.

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

Как будем решать?

- One vs. One

Linear models. Classification.

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

Как будем решать?

- One vs. One
- One vs. All

Linear models. Classification.

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

Как будем решать?

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

Linear models. Classification.

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

Как будем решать?

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

- Линейная регрессия

Linear models. Classification.

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

Как будем решать?

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

- Линейная регрессия
- Логистическая регрессия

Linear models. Classification.

Перейдем к другим обозначениям. Пусть линейная модель записывается:

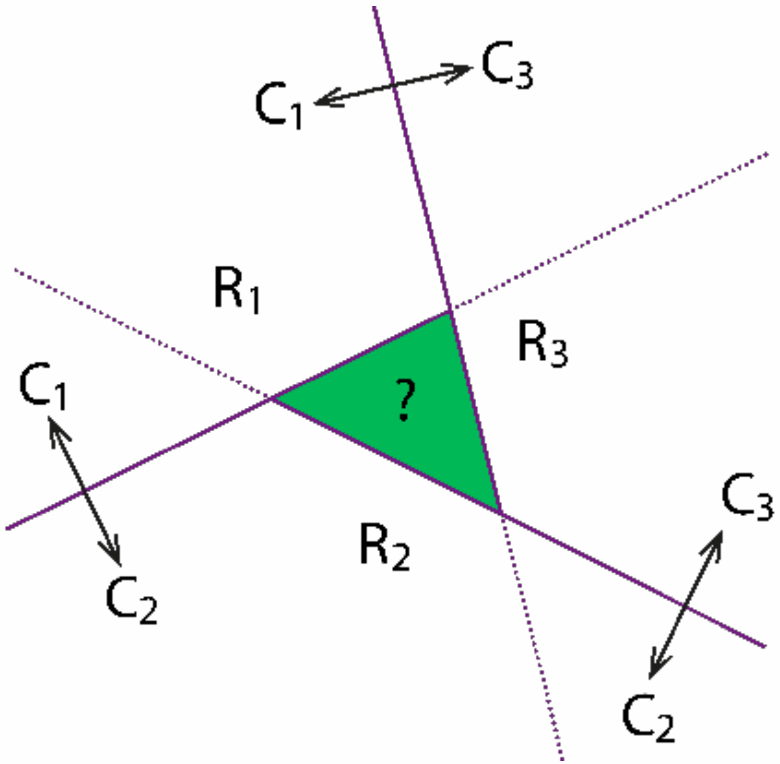
$$\mathbf{y}(\mathbf{x}) = \mathbf{x}'\mathbf{w} + \mathbf{w_0}$$

Пусть объект, описываемый вектором x относится к классу 1, если $y(x) \geq 0$ и к классу 2 в противном случае.

Linear models. One vs. One.

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

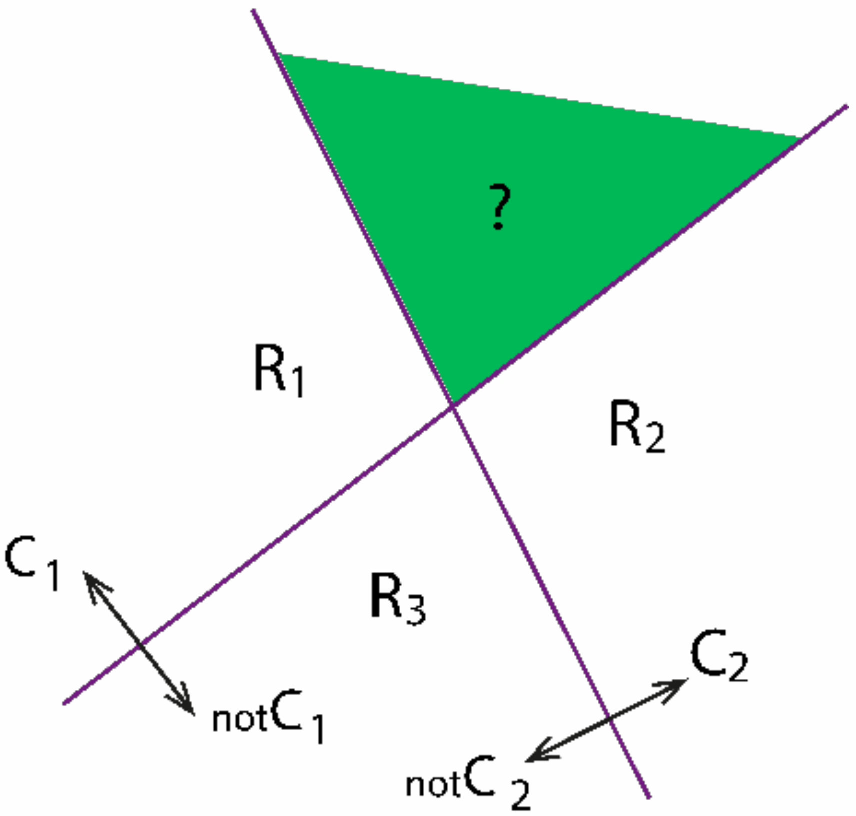
Проблема:



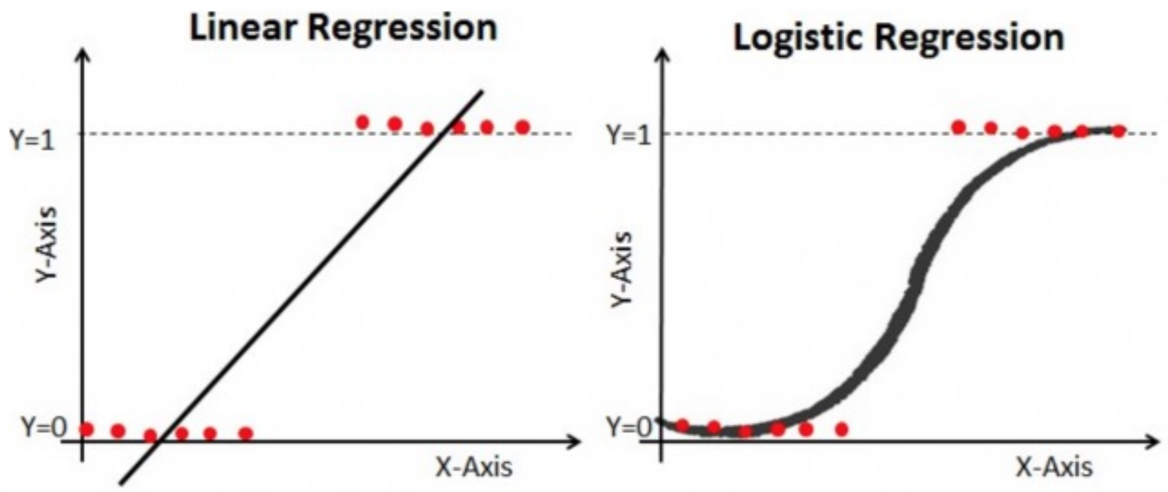
Linear models. One vs. All.

Строим K-1 бинарных классификаторов.

Проблема:



Logistic Regression



https://medium.com/@ODSC/logistic-regression-with-python-ede39f8573c7

Базируется на нелинейном преобразовании.
Позволяет учесть нормирование на единицу и неотрицательность вероятности.

Logistic Regression

Это осуществляется вводом логит функции или её преобразования – сигмоиды.

Logistic Regression

Это осуществляется вводом логит функции или её преобразования – сигмоиды:

$$P(y = 1|\mathbf{X})$$

Logistic Regression

Это осуществляется вводом логит функции или её преобразования – сигмоиды:

$$P(y = 1|\mathbf{X})$$

$$\ln \frac{P(y_i = 1|\mathbf{X}_i)}{P(y_i = 0|\mathbf{X}_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Logistic Regression

Это осуществляется вводом логит функции или её преобразования – сигмоиды:

$$P(y = 1|\mathbf{X})$$

$$\ln \frac{P(y_i = 1|\mathbf{X}_i)}{P(y_i = 0|\mathbf{X}_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

$$z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

$$P(y_i = 1|\mathbf{X}_i) = \frac{1}{1 + \exp(-z)} + \epsilon_i = \sigma(z_i) + \epsilon_i$$

Logistic Regression

Для каждого наблюдения получаем вероятность увидеть y=1 при данных характеристиках.

Logistic Regression

Для каждого наблюдения получаем вероятность увидеть y=1 при данных характеристиках.

Можем записать функцию правдоподобия:

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ P(y_i = 1|X_i)^{y_i} \left(1 - P(y_i = 1|X_i)^{1-y_i}\right) \right\}$$

Максимизируя её по нашим параметрам бета и находя оценки мы находим оценки вероятностей исходов.

Logistic Regression

Статистический пакет представляет результаты оценки логистической регрессии в следующем виде:

```
         Current function value: 0.525192
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                32561
Model:                          Logit   Df Residuals:                    32559
Method:                           MLE   Df Model:                            1
Date:                Tue, 13 Feb 2018   Pseudo R-squ.:                 0.04859
Time:                        21:24:30   Log-Likelihood:                -17101.
converged:                       True   LL-Null:                       -17974.
                                        LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         -2.7440      0.043    -64.211      0.000        -2.828    -2.660
age            0.0395      0.001     40.862      0.000         0.038     0.041
==============================================================================
```

https://sweetcode.io/easy-scikit-logistic-regression/

В этой задаче классификация проводилась для события “человек зарабатывает более 50 тыс. долларов” по одной переменной – возрасту.

Давайте посмотрим как интерпретировать результаты.

Logistic Regression

Коэффициент при age положительный, значимый на любом разумном уровне значимости и равен 0.0395.

Что происходит при увеличении возраста на 1 год?
В логистической регрессии проще всего это иллюстрировать “отношением шансов” (odds ratio).

$$\frac{\frac{P(y=1|age+1)}{P(y=0|age+1)}}{\frac{P(y=1|age)}{P(y=0|age)}} = \frac{\exp\{\alpha + \beta(age+1)\}}{\exp\{\alpha + \beta age\}} = \exp\{\beta\}$$

То есть, при увеличении возраста на 1 год шансы получать более 50к возрастают в 1.04 раза.
Вероятность также можно получить, подставив значения коэффициентов в формулы, данные вначале.

Logistic Regression

Мы можем использовать либо оценку вероятности ($\hat{p}$), либо предсказанный класс.

Но мы имеем только оценку вероятности, как же получить предсказание?

Logistic Regression

Мы можем использовать либо оценку вероятности ($\hat{p}$), либо предсказанный класс.

Но мы имеем только оценку вероятности, как же получить предсказание?

Для этого необходимо выбрать порог (threshold), такой, что:

$$\hat{y} \in \{0, 1\}$$

$$\hat{y} = \begin{cases} 0, \hat{p} \leq threshold \\ 1, \hat{p} > threshold \end{cases}$$

Linear models. Multiclass.

Рассмотрим multiclass logistic regression.

Предположим $y \in \{1, 2, ..., K\}$.

Linear models. Multiclass.

Рассмотрим multiclass logistic regression.

Предположим $y \in \{1, 2, ..., K\}$.

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

Тогда, по аналогии с предыдущим запишем модель:

$$\ln \frac{P(y=1|\mathbf{x})}{P(y=K|\mathbf{x})} = w_{10} + w_1'\mathbf{x}$$

$$\ln \frac{P(y=2|\mathbf{x})}{P(y=K|\mathbf{x})} = w_{20} + w_2'\mathbf{x}$$

$$\dots$$

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

Linear models. Multiclass.

Отсюда легко получить:

$$P(y = i|\mathbf{x}) = \exp\left(w_{i0} + w_i'\mathbf{x}\right)P(y = K|\mathbf{x})$$

Linear models. Multiclass.

Отсюда легко получить:

$$P(y = i|\mathbf{x}) = \exp\left(w_{i0} + w_i'\mathbf{x}\right)P(y = K|\mathbf{x})$$

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

Linear models. Multiclass.

Отсюда легко получить:

$$P(y=i|\mathbf{x}) = \exp\left(w_{i0} + w_i'\mathbf{x}\right)P(y=K|\mathbf{x})$$

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

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

Linear models. Multiclass.

Отсюда легко получить:

$$P(y=i|\mathbf{x}) = \exp\left(w_{i0} + w_i'\mathbf{x}\right)P(y=K|\mathbf{x})$$

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

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

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