MACHINE
LEARNING

Ниатшин Булат
Data scientist, Рекламные технологии

Задача классификации. Логистическая регрессия, KNN

10.10.2019

MACHINE
LEARNING

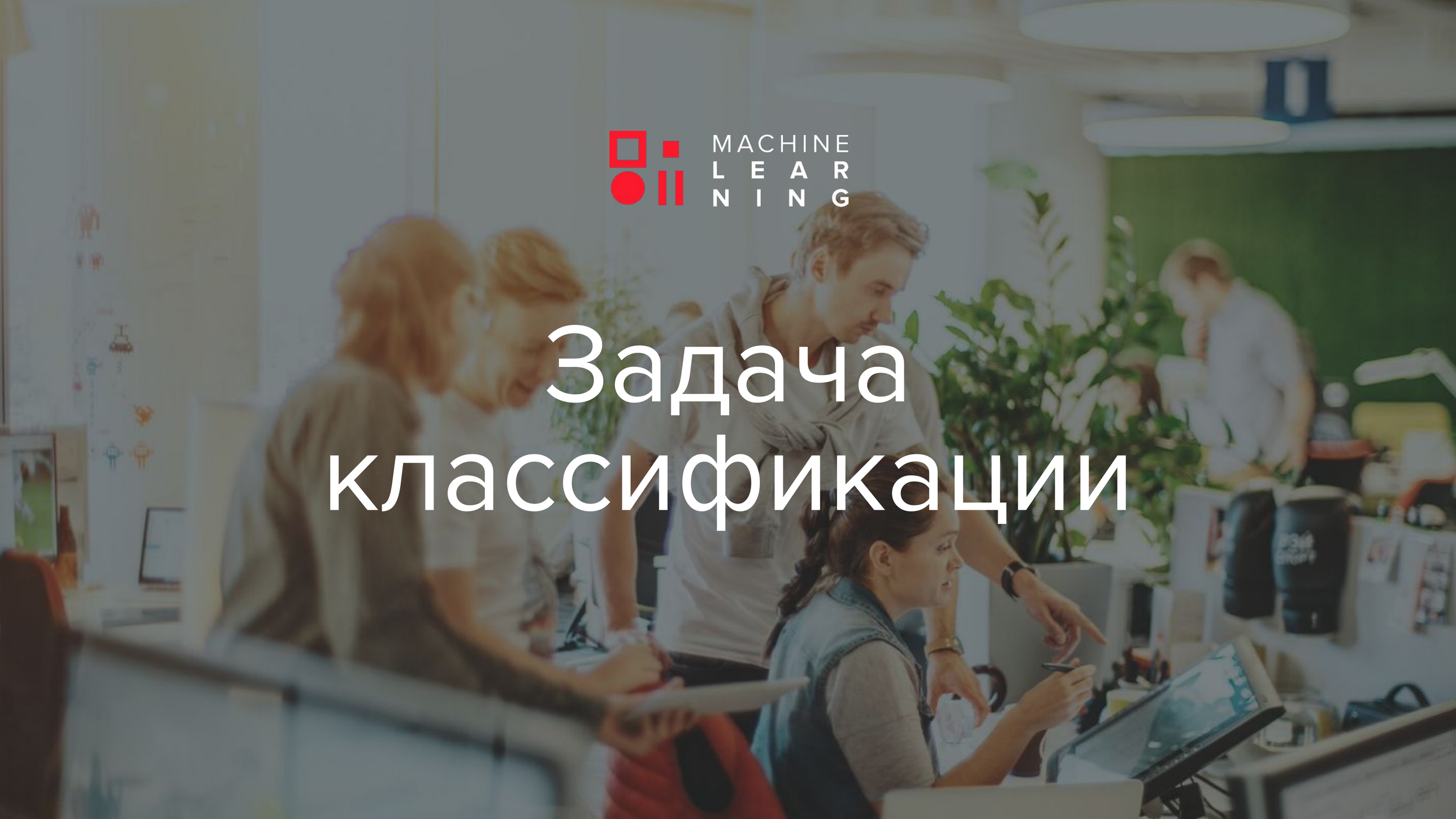
О себе

- Рекламные технологии, Big Data
- Ранжирование и сегментация пользователей
- Разработка ML инфраструктуры

Данные: название источника данных, 2016

План занятия

- Обзор и постановка задачи классификации
- Линейный классификатор
- Логистическая регрессия, метод максимального правдоподия.
- KNN
- Семинар



Задача классификации

Жизненный пример

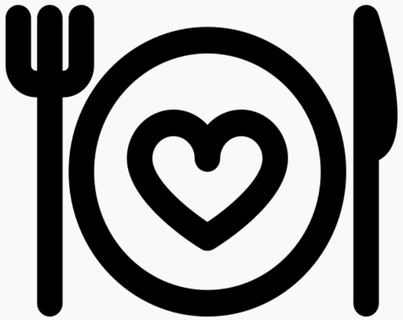
MACHINE LEARNING

Выходить ли мне сегодня из дома?

+1

Вы свободны сейчас

+2



Вы хотите есть

-3

Вы хотите спать

+5

Вы хотите тусить

Жизненный пример

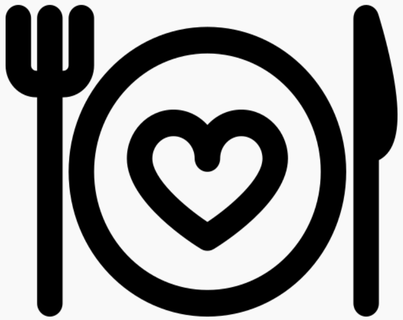
MACHINE LEARNING

Выходить ли мне сегодня из дома?

+1

Вы свободны сейчас

+2



Вы хотите есть

-3

Вы хотите спать

+5

Вы хотите тусить

Сумма >= **1** - значит выходим :)

Более серьезный пример

MACHINE LEARNING

Банковский скоринг

Дать или не дать человеку кредит и на каком основании

ПОКАЗАТЕЛЬ	ДИАПАЗОН ЗНАЧЕНИЙ
Возраст заемщика	До 35 лет
	От 35 до 45 лет
	От 45 и старше
Образование	Высшее
	Среднее специальное
	Среднее
Состоит ли в браке	Да
	Нет
Наличие кредита в прошлом	Да
	Нет
Стаж работы	До 1 года
	От 1 до 3 лет
	От 3 до 6 лет
	Свыше 6 лет
Наличие автомобиля	Да
	Нет

Более серьезный пример

MACHINE LEARNING

Банковский скоринг

Суммируем скоринг-балл, если значение выше порога, то выдаем кредит.

ПОКАЗАТЕЛЬ	ДИАПАЗОН ЗНАЧЕНИЙ	СКОРИНГ-БАЛЛ
Возраст заемщика	До 35 лет	7,60
	От 35 до 45 лет	29,68
	От 45 и старше	35,87
Образование	Высшее	29,82
	Среднее специальное	20,85
	Среднее	22,71
Состоит ли в браке	Да	29,46
	Нет	9,38
Наличие кредита в прошлом	Да	40,55
	Нет	13,91
Стаж работы	До 1 года	15,00
	От 1 до 3 лет	18,14
	От 3 до 6 лет	19,85
	Свыше 6 лет	23,74
Наличие автомобиля	Да	51,69
	Нет	15,93

Почему так нельзя дальше

MACHINE LEARNING

Подбор весов и порога

- Сложно делать вручную
- Требуется экспертиза в области
- Требуется проверка на данных (эксперт может ошибиться и что-то не учесть).

Выход

Автоматизируем подбор параметров. Будем использовать методы численной оптимизации для решения задачи.

Данные: название источника данных, 2016

Формальная постановка задачи

MACHINE LEARNING

Рассмотрим случай бинарной классификации

Пусть $\mathbb{X} = \mathbb{R}^n$ - пространство объектов

$\mathbb{Y} = \{+1, -1\}$ - множество допустимых ответов

$X = \{(x_i, y_i)\}_{i=1}^{l}$ - обучающая выборка

$a(x) = sign(\langle \omega, x \rangle)$ - линейный классификатор, полученный в результате

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i] = \frac{1}{\ell} \sum_{i=1}^{\ell} [sign(\langle w, x_i \rangle) \neq y_i] \to \min_w$$

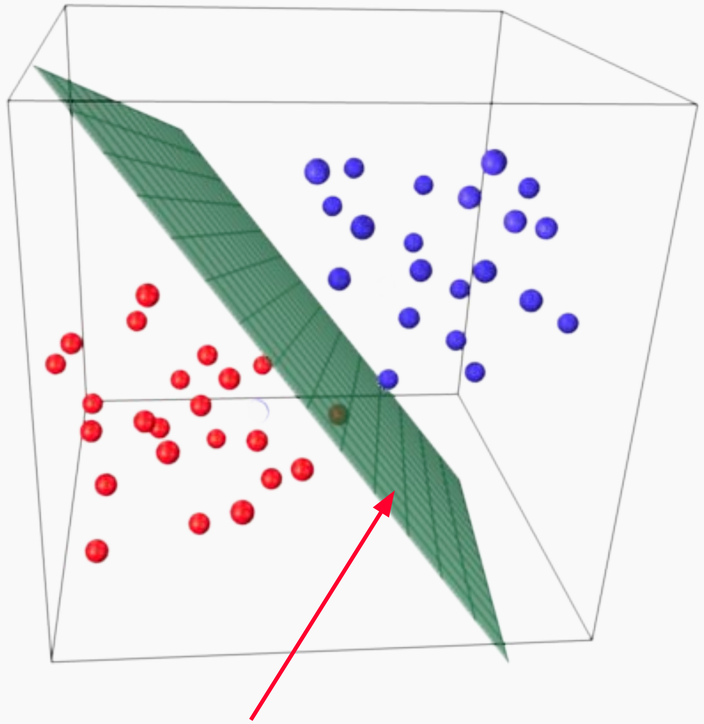
Линейный классификатор

MACHINE LEARNING

$$\alpha(x) = \begin{cases} 1, & \text{if } f(x) > 0 \\ -1, & \text{if } f(x) \leq 0 \end{cases}$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_k x_k$$
$$f(x) = w_0 + \langle w, x \rangle$$
$$f(x) = \langle w, x \rangle$$

Интерпретация: строим разделяющую плоскость



$$f(x) = \langle w, x \rangle = 0$$

Логистическая регрессия

MACHINE LEARNING

- Частный случай линейного классификатора
- Умеет предсказывать вероятность принадлежности к позитивному классу
- Один из самых распространенных классических алгоритмов

Для чего нужна вероятность помимо метки класса?

Логистическая регрессия

MACHINE LEARNING

- Частный случай линейного классификатора
- Умеет предсказывать вероятность принадлежности к позитивному классу
- Один из самых распространенных классических алгоритмов

Для чего нужна вероятность помимо метки класса? - Банковский скоринг, формирование новостной ленты и т.д.

Логистическая регрессия

MACHINE LEARNING

Что мы умеем на данный момент:

- Строить линейный прогноз при помощи МНК: $\hat{y(x)} = \omega^T x \in \mathbb{R}$

Что хотим получить:

- Преобразовать прогноз в вероятность, принимающее значение в интервале [0, 1]

Логистическая регрессия

MACHINE LEARNING

Что мы умеем на данный момент:

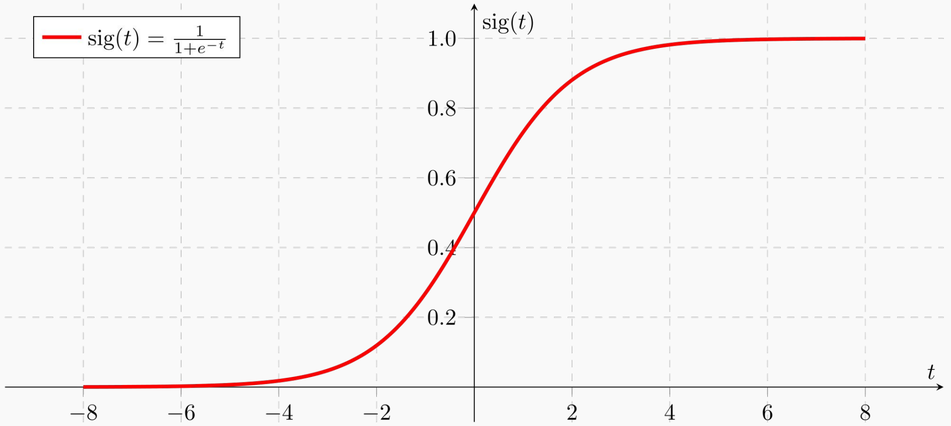
- Строить линейный прогноз при помощи МНК: $y(\hat{x}) = \omega^T x \in \mathbb{R}$

Что хотим получить:

- Преобразовать прогноз в вероятность, принимающее значение в интервале [0, 1]

Выход? Сигмоидная функция:

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



Вывод алгоритма

MACHINE LEARNING

Вероятность происхождения события X - $P(X)$

Отношение вероятностей - $OR(X) = \frac{P(X)}{1 - P(X)}$

Вычислим логарифм отношения вероятностей - $log(OR(X)) \in \mathbb{R}$

Шаг 1. Вычисляем $\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_k x_k$

Шаг 2. Вычисляем логарифм отношения вер-ей $log(OR) = \omega^T x$

Шаг 3. Вычисляем искомую вероятность:

$$p_+ = \frac{OR_+}{1 + OR_+} = \frac{1}{1 + exp^{-\omega^T x}} = \sigma(\omega^T x)$$

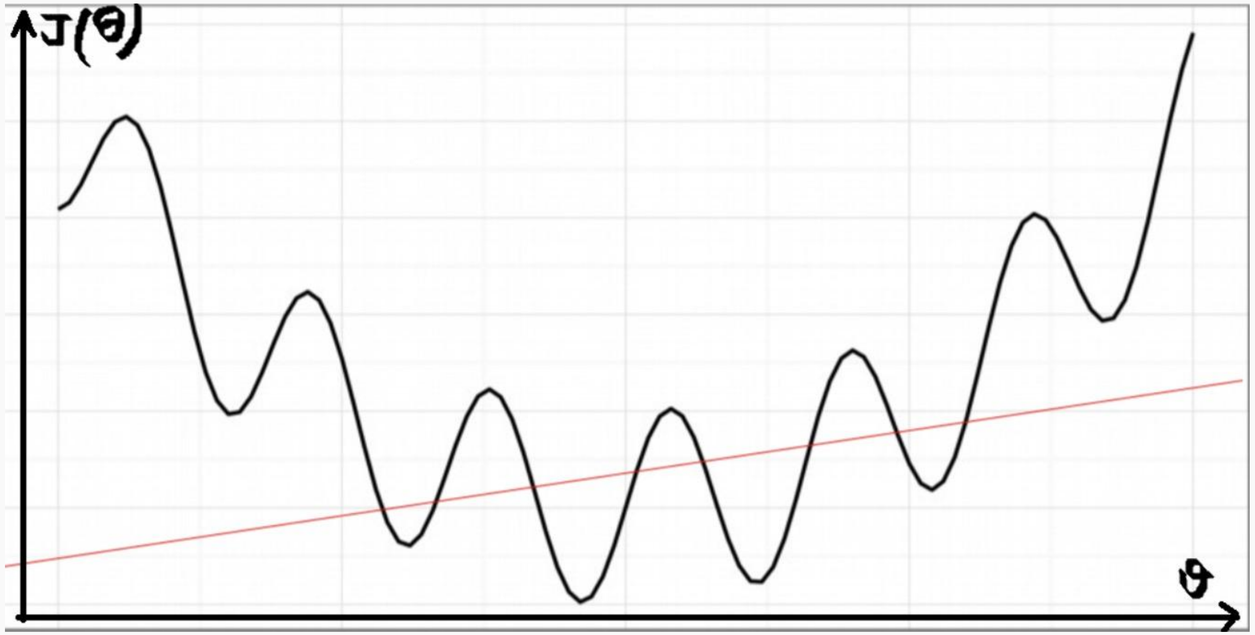
Loss function

MACHINE LEARNING

Для линейной регрессии:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$



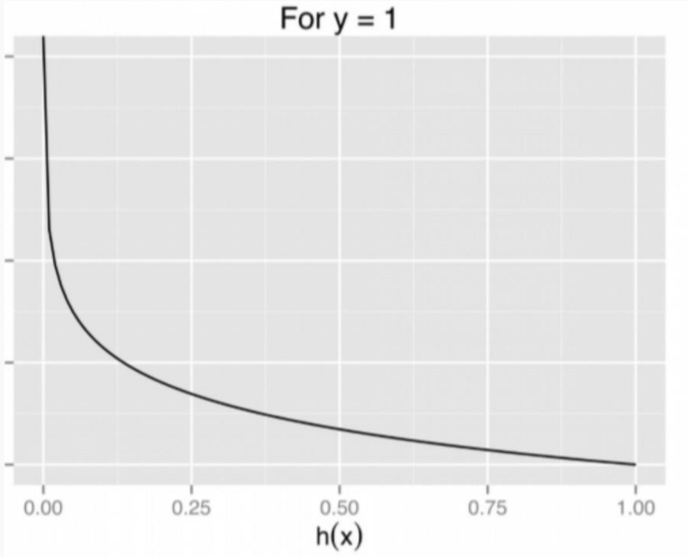
Проблема - невыпуклая функция в случае сигмоиды

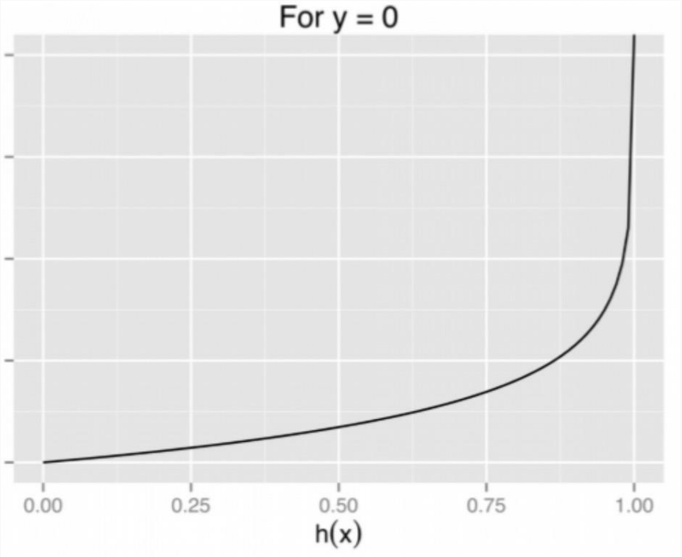
Loss function

MACHINE LEARNING

Рассмотрим функцию следующего вида:

$$L(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)), & y = 1 \\ -log(1 - h_\theta(x)), & y = 0 \end{cases}$$





Loss function

MACHINE LEARNING

$$J(h_\theta(x), y) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)})(1 - log h_\theta(x^{(i)}))]$$

Регуляризация

MACHINE LEARNING

$$J(X, y, \theta) = J(X, y) + \frac{1}{C}||\theta||^2$$

- C - обратный коэффициент регуляризации (как в sklearn)
- Больше значение C - больше “сложность модели”
- Малые значения C - недообученность модели
- Гиперпараметр, который необходимо подбирать на кросс-валидации

kNN

MACHINE LEARNING

k Nearest Neighbours (k Ближайших соседей):

- Один из самых популярных методов для задач классификации
- Используется также для регрессии
- Хорошо изученный подход, имеющий сильную теорию под собой

kNN

MACHINE LEARNING

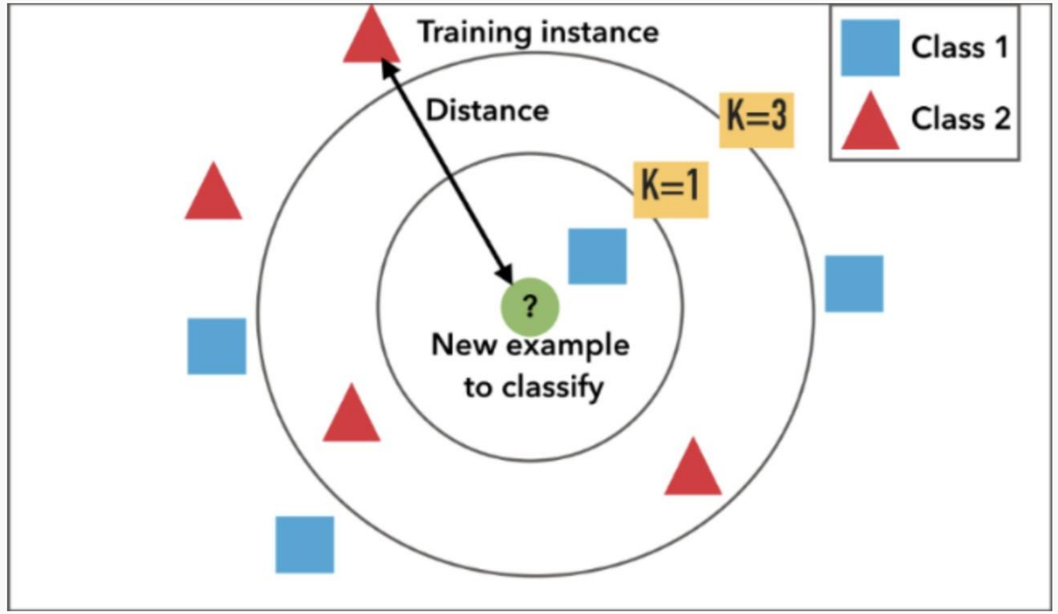
Имеется **обучающая выборка:** *X, y*

Метрика схожести объектов: *q*

Алгоритм:

Пусть на входе объект test выборки. Для него:

1. Вычисляем расстояния до каждого объекта train
2. Выбираем k наиближайших
3. Тестовому объекту присваиваем метку класса, которая присутствует больше всех среди этих k объектов

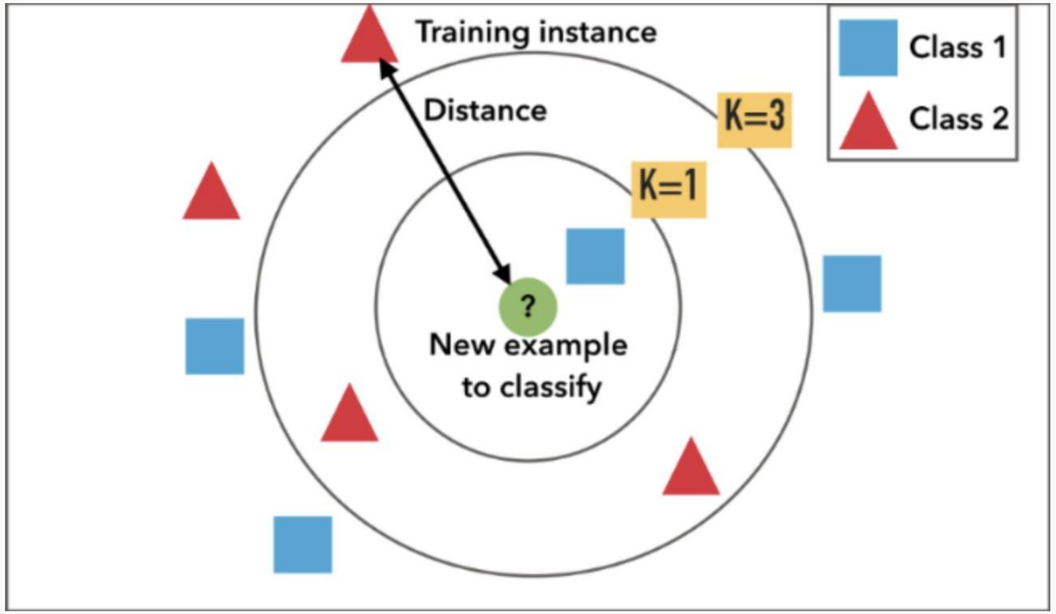


kNN

MACHINE LEARNING

Модификации:

1. Каждому объекту из k наиближайших присваиваем вес - 1 / distance
2. Рассматриваем объекты только в пределах заданного радиуса R
3. Прореживание большой выборки (выбрасывание неинформативных объектов)
4. Большое количество различных вариантов реализаций approximate nearest neighbour (к примеру, библиотека annoy от Spotify)

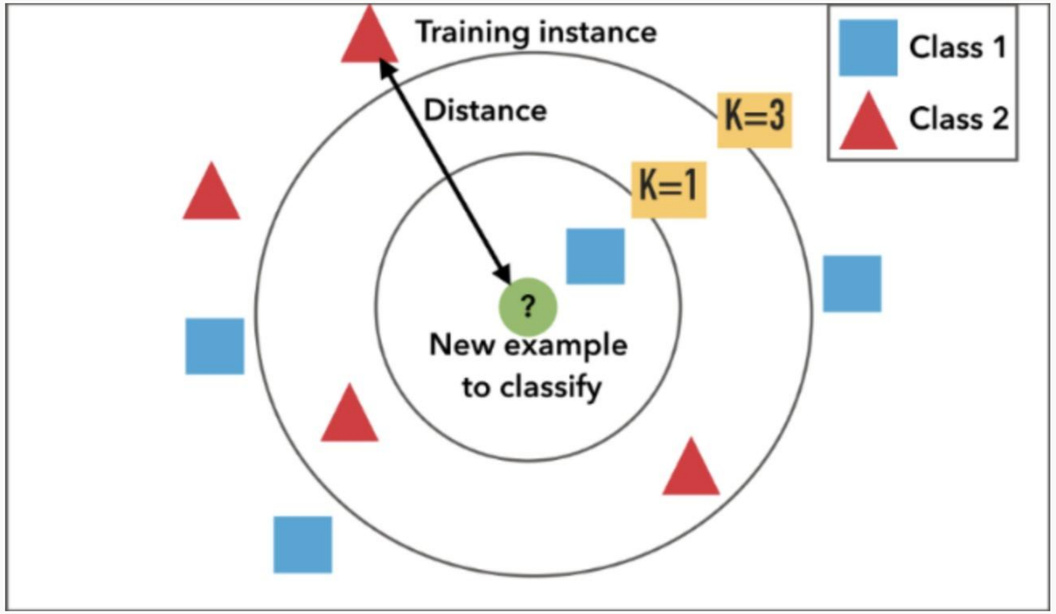


kNN

MACHINE LEARNING

Применение в реальных задачах:

1. Baseline для исследовательских задач
2. Рекомендательные системы - поиск похожих товаров, людей и т.д.
3. Стекинг/Блендинг
4. Мета-признаки в Kaggle соревнованиях



Метрики качества классификации

MACHINE LEARNING

1. Accuracy
2. Precision
3. Recall
4. ROC-AUC
5. F-score

Метрики качества классификации

MACHINE LEARNING

Accuracy:

Подсчитываем долю правильно предсказанных объектов.

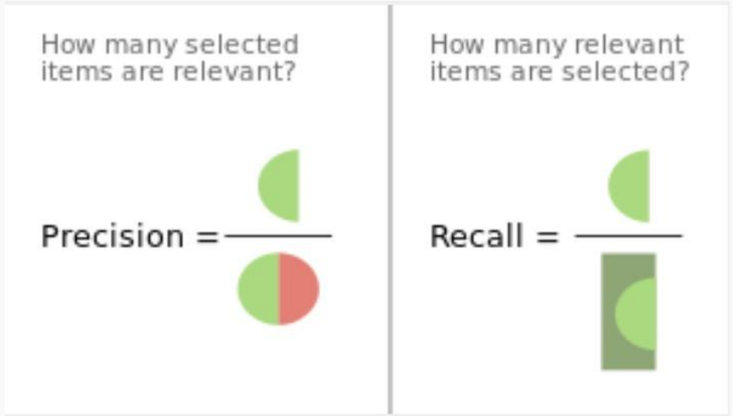
```
1. import numpy as np
2. target = np.array([1, 3, 2, 2, 3, 4, 1, 2])
3. pred = np.array([1, 3, 1, 2, 3, 2, 1, 2])
4. print(np.equal(target, pred).sum())
5. print(np.equal(target, pred).sum() / float(target.shape[0])
```

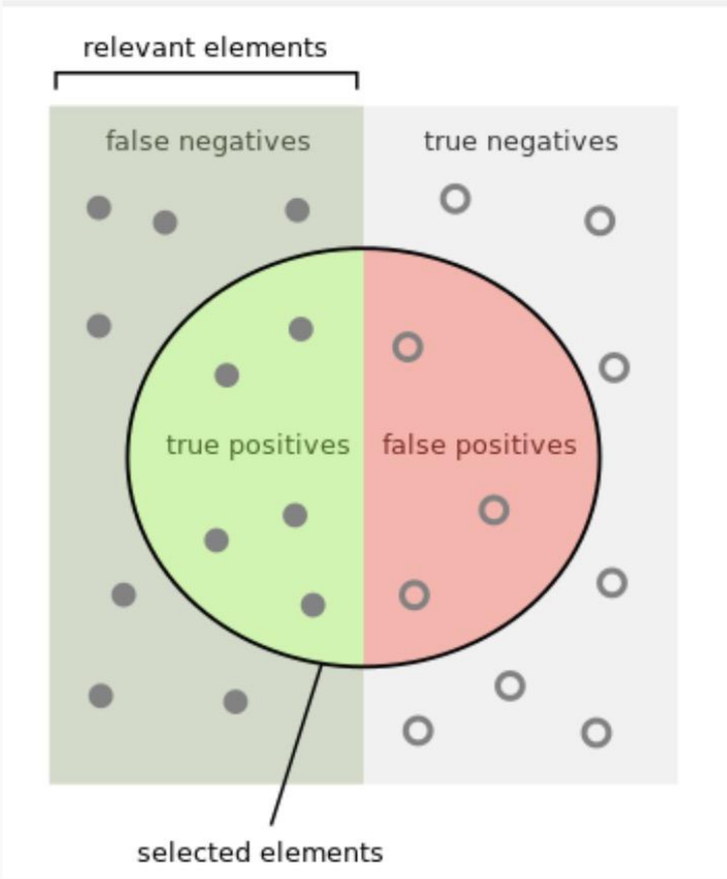
Результат:
6
0.75

Метрики качества классификации

MACHINE LEARNING

Precision/Recall - бинарная классификация:





Метрики качества классификации

MACHINE LEARNING

Precision/Recall - бинарная классификация:

	Предсказали True	Предсказали False
Ожидали True	True Positive (tp)	False Negative (fn)
Ожидали False	False Positive (fp)	True Negative (tn)

Полнота	Recall = tp / (tp + fn)	Какую часть объектов класса 1 мы нашли?
Точность	Precision = tp / (tp + fp)	Какая часть из найденных объектов класса 1 действительно ими является?

Метрики качества классификации

MACHINE LEARNING

Precision/Recall:

1. target = np.array([0, 1, 1, 0, 1, 1])
2. pred = np.array([1, 1, 1, 1, 1, 1])
3. print(precision(target, pred))
4. print(recall(target, pred))

Результат:

Recall - 1 = 4 / (4 + 0) Какую часть из объектов класса 1 мы нашли?

Precision - ⅔ = 4 / (4 + 2) Какая часть из найденных объектов класса 1 действительно 1?

Метрики качества классификации

MACHINE LEARNING

Кросс-валидация

Процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецендентам.

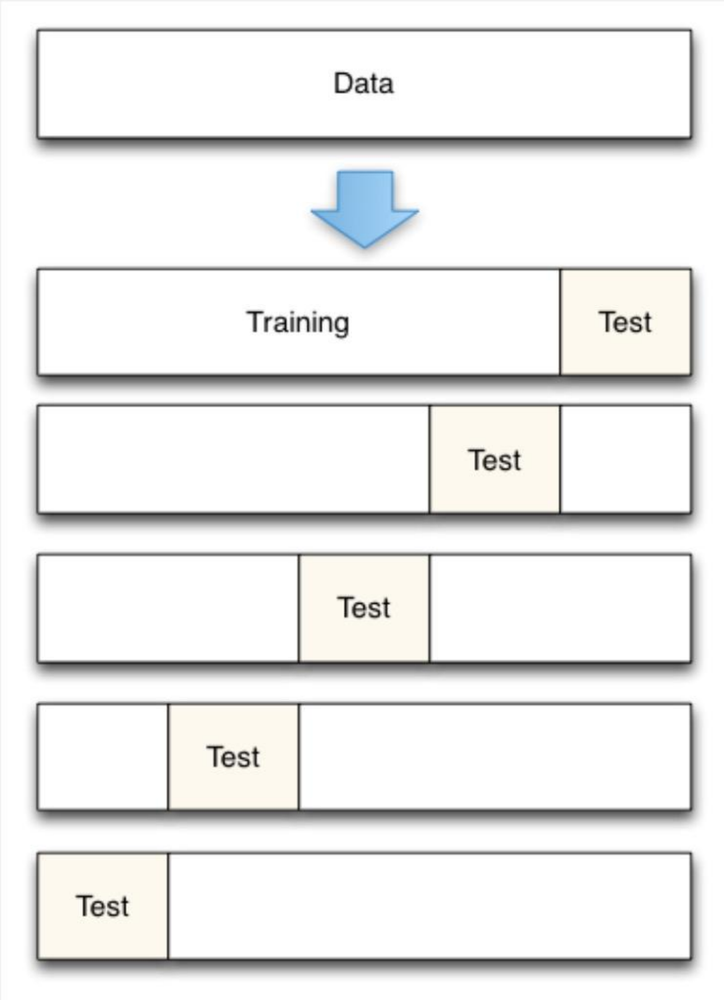
Фиксируется некоторое множество разбиений исходной выборки на две подвыборки:

1. Обучающую
2. Контрольную

Метрики качества классификации

MACHINE LEARNING

Кросс-валидация



Спасибо за внимание!

Ниатшин Булат

+7 999 965-61-63
b.niatshin@corp.mail.ru

MACHINE LEARNING