ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

Лекция №1

Осень 2024/2025

Формат курса

Оценка за курс:

- устный ответ по программе курса
  - 2 вопроса из программы (на зачёте)
- 3 дополнительных балла
  - работа на семинарах
  - решение дополнительных заданий

- Лекционные занятия: онлайн/оффлайн + запись
- Семинарские занятия (по группам): онлайн/оффлайн + запись
- Домашние задания с фиксированным дедлайном
  - Проверка семинаристами по группам

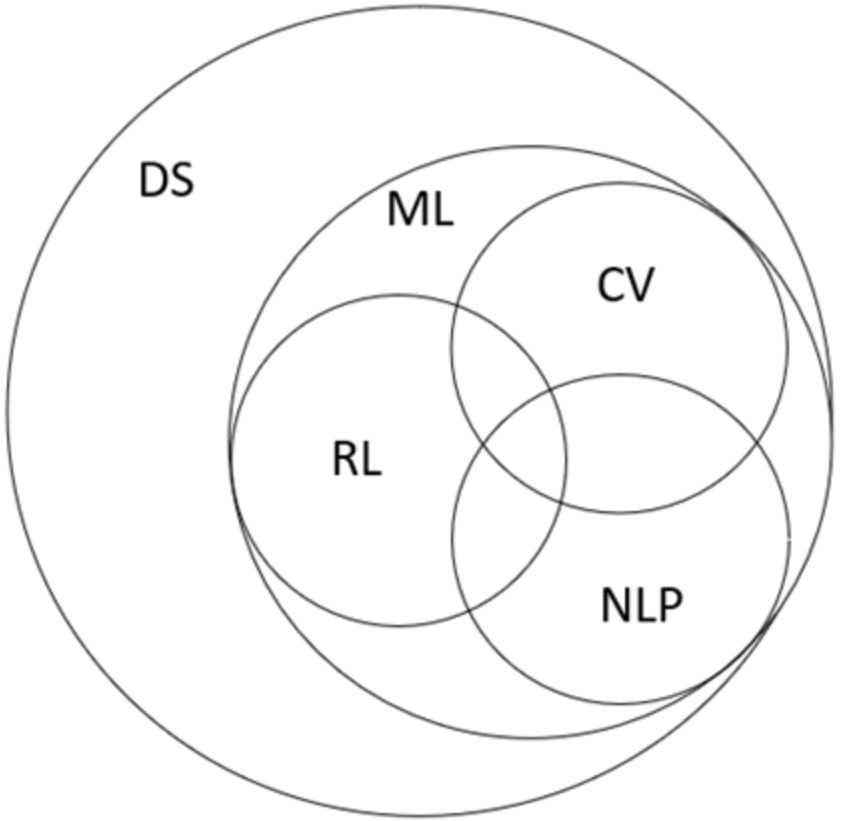
Программа курса

1. Naive Bayes, kNN
2. Линейные модели
3. Логистическая регрессия
4. SVM, PCA
5. BVD, k
6. Деревья решений. Методы ансамблирования моделей
7. Градиентный бустинг
8. Введение в нейронные сети
9. Методы кластеризации и понижения размерности
10. Неградиентная оптимизация
11. Задачи ранжирования и матчинга

Введение

Три основных области исследований в ML (Machine learning)

1. CV (Computer Vision)
2. NLP (Natural Language Processing)
3. RL (Reinforcement Learning)



Коротко о задачах в ML

Решим задачу

Сколько минут в 3 часах?

Коротко о задачах в ML

Решим задачу

Сколько минут в 3 часах?

$f(x) = 60 * x$
$f(3) = 60 * 3 = 180$

Решим другую задачу

Мальчик на санках едет с горки. Масса мальчика вместе с санками составляет 40 кг, угол наклона горы 30°. Найдите ускорение, с которым съезжает мальчик, если коэффициент трения скольжения равен 0,2.

Решим другую задачу

Мальчик на санках едет с горки. Масса мальчика вместе с санками составляет 40 кг, угол наклона горы 30°. Найдите ускорение, с которым съезжает мальчик, если коэффициент трения скольжения равен 0,2.

Дано:

$m = 40$ кг

$\alpha = 30°$

$\mu = 0{,}2$

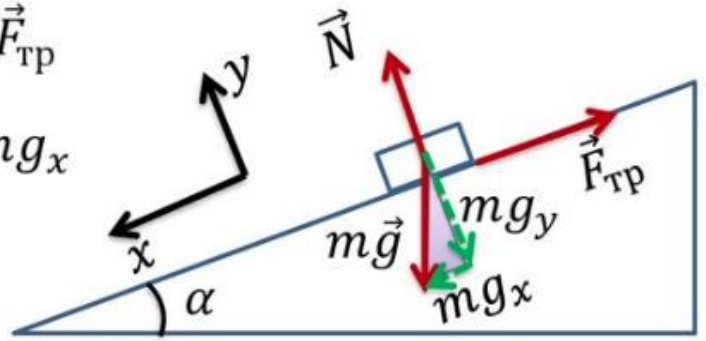
$a - ?$

$$m\vec{a} = \vec{N} + m\vec{g} + \vec{F}_{тр}$$

$$X: ma = -F_{тр} + mg_x$$

$$Y: 0 = N - mg_y$$

$$N = mg_y$$



$$F_{тр} = \mu N = \mu mg_y$$

$$ma = mg_x - \mu mg_y$$

$$mg_x = mg\sin\alpha$$

$$mg_y = mg\cos\alpha$$

А что если?

- система сложнее?
- процесс сложнее?
- мы не имеем представления, как он устроен?
- мы не понимаем, как параметры внутри влияют друг на друга?

Просто "потрясающий" фильм. Про него даже сказать нечего. Не то, чтобы он мне не понравился, он просто никакой.

Проанализируйте окрас текста

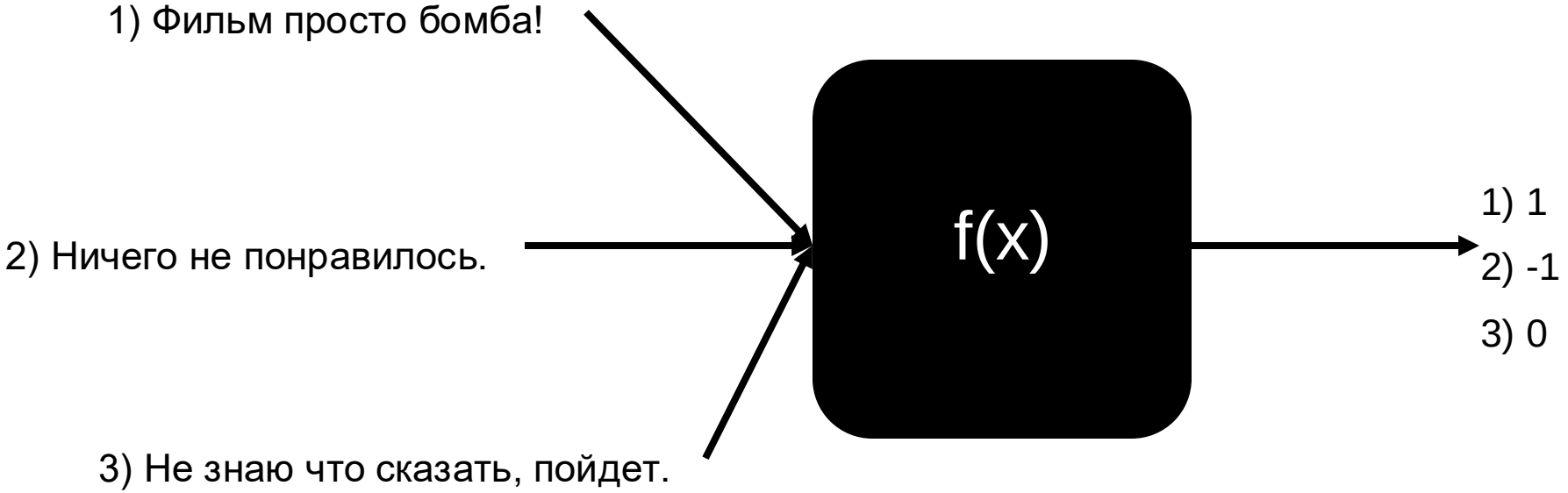
позитивный / нейтральный / негативный

Просто потрясающий фильм. У меня даже слов нет. Не шедевр десятилетия, но открытие этого года - точно.

Предположим, что...

Есть некоторая "магическая коробка" которая:

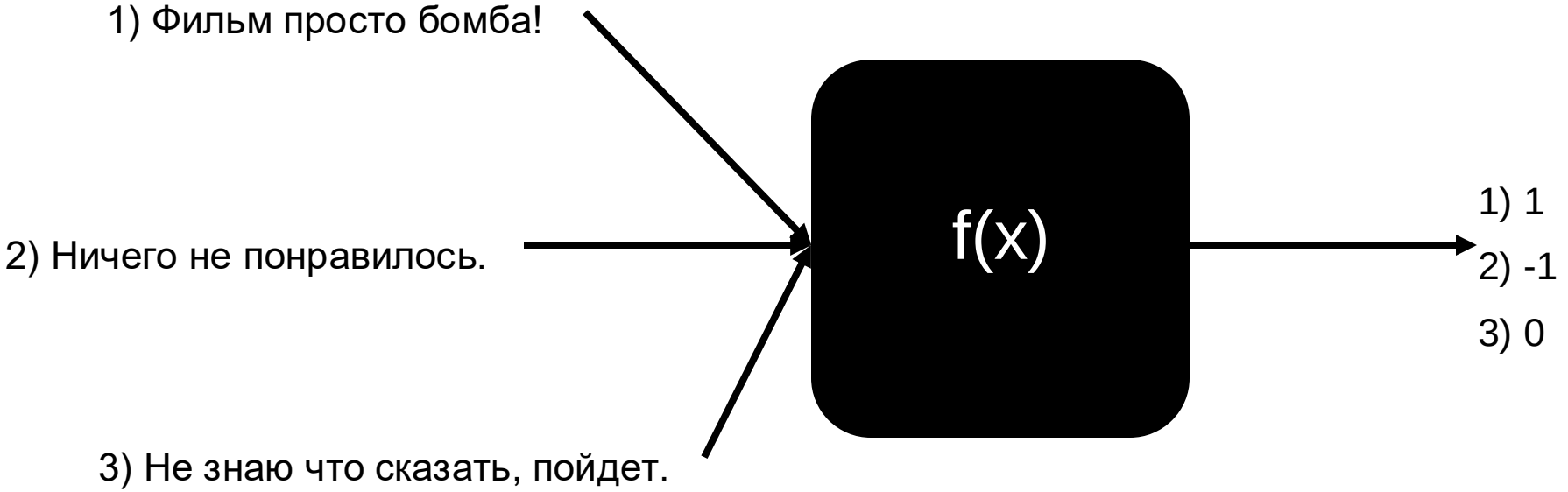
- получает текст
- выдает 1, 0, -1



Предположим, что...

Есть некоторая "магическая коробка" которая:

- получает текст
- выдает 1, 0, -1



... вот только такую f(x) мы не можем придумать

01

Основные понятия и термины

Основные термины

Пускай есть задача открыть новый ресторан, есть несколько вариантов размещения, какой из них принесет максимальную прибыль?

- x - объект для которого делаем предсказание
  конкретное расположение ресторана
- $\mathbb{X}$ - пространство объектов
  все возможные расположения ресторанов
- y - ответ, целевая переменная / метка, target (что предсказываем)
  прибыль в течение первого года работы
- $\mathbb{Y}$ - пространство ответов (все возможные значения ответа)
  все вещественные числа

Основные термины

Обучающая выборка

- мы не разбираемся в области той задачи, которую решаем
- у нас есть множество объектов с известными ответами
- $X = (x^i, y^i)_{i=1}^{l}$ - обучающая выборка
- l - размер выборки

Признаки

- каждый объект как либо описан в числовом виде
- признаки (features) - числовые характеристики объекта
- $x = (x^i, \dots x^d)$ - признаковое описание объекта (x - вектор)
- d - количество признаков

Основные термины

Обучающая выборка

- мы не разбираемся в области той задачи, которую решаем
- у нас есть множество объектов с известными ответами
- $X = (x^i, y^i)_{i=1}^{l}$ - обучающая выборка
- l - размер выборки

  Пример: существующие данные по расположению и выручке

Признаки

- каждый объект как либо описан в числовом виде
- признаки (features) - числовые характеристики объекта
- $x = (x^i, \dots x^d)$ - признаковое описание объекта (x - вектор)
- d - количество признаков

Пример: удаленность от дороги / метро, стоимость квадратного метра жилья поблизости, количество бизнес-центров вокруг, средняя проходимость за день

Основные термины

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Основные термины

Алгоритм

- a(x) - алгоритм, модель - функция, на входе принимающая числа и выдающая ответ для любого объекта
- отображает $\mathbb{X}$ в $\mathbb{Y}$
- линейная модель: $a(x) = w^0 + w^1x^1 + \ldots + w^dx^d$

Функция потерь

- позволяет понять полезность алгоритма, a(x) = 0 нам не поможет
- функция потерь - мера корректности ответа алгоритма
- что лучше - предсказать больше или меньше? предсказали 100.000, а на деле 90.000 - это хорошо или плохо?

Основные термины

Алгоритм

- a(x) - алгоритм, модель - функция, на входе принимающая числа и выдающая ответ для любого объекта
- отображает $\mathbb{X}$ в $\mathbb{Y}$
- линейная модель: $a(x) = w^0 + w^1 x^1 + ... + w^d x^d$

Пример: a(x) = 90.000 + 10.000 * (кол-во бизнес-центров вокруг) - 2.000 * (удаленность от метро)

Функция потерь

- позволяет понять полезность алгоритма, a(x) = 0 нам не поможет
- функция потерь - мера корректности ответа алгоритма
- что лучше - предсказать больше или меньше? предсказали 100.000, а на деле 90.000 - это хорошо или плохо?

Пример: квадратичное отклонение: $(a(x) - y)^2$

Основные термины

Функционал ошибки

- функционал ошибки , метрика качества - мера качества работы алгоритма на выборке
- чем меньше, тем лучше
- выбирается исходя из бизнес-требований конкретной задачи

Обучение алгоритма

- производится с использованием обучающей выборки и функционала ошибки
- есть некоторое семейство алгоритмов $\mathcal{A}$
- из семейства выбираем алгоритм
- обучение - поиск оптимального алгоритма с точки зрения функционала ошибки
- с точки зрения функционала ошибки = минимизируем ошибку

Основные термины

Функционал ошибки

- функционал ошибки , метрика качества - мера качества работы алгоритма на выборке
- чем меньше, тем лучше
- выбирается исходя из бизнес-требований конкретной задачи

Пример: Среднеквадратичная ошибка (MSE): $\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2$

Обучение алгоритма

- производится с использованием обучающей выборки и функционала ошибки
- есть некоторое семейство алгоритмов $\mathcal{A}$
- из семейства выбираем алгоритм
- обучение - поиск оптимального алгоритма с точки зрения функционала ошибки
- с точки зрения функционала ошибки = минимизируем ошибку

Пример: линейные модели; $\mathcal{A} = \{w_0 + w_1x_1 + \cdots + w_dx_d \mid w_0, w_1, \ldots, w_d \in \mathbb{R}\}$

тогда обучение: $a(x) = \arg\min_{a\in\mathcal{A}} Q(a, X)$

02

Какие бывают признаки?

Признаки

Бинарные

- $D_j = \{0,1\}$
- Ресторан находится в бизнес-центре?
- Пол клиента
- Компания вышла на IPO?

Вещественные

- $D_j = \mathbb{R}$
- Средний доход жильцов дома
- Средняя выручка компании за последние 5 лет
- Возраст клиента

Признаки

Категориальные

- D_j = неупорядоченное множество
- Район / город расположения квартиры
- Цвет глаз / волос человека
- Статус работы человека

Порядковые

- D_j = упорядоченное множество
- Воинское звание
- Должность на работе (может быть и не упорядоченным)
- Тип населенного пункта

Признаки

Категориальные

- D_j = неупорядоченное множество
- Район / город расположения квартиры
- Цвет глаз / волос человека
- Статус работы человека

Вещественные

- D_j = упорядоченное множество
- Воинское звание
- Должность на работе (может быть и не упорядоченным)
- Тип населенного пункта

А как с ними работать? Помним: компьютер оперирует только числами

Машинное обучение это…

- Нахождение зависимостей из конечного набора примеров.

- Написание алгоритма, способного решать задачи, которые не явно запрограммированы (последовательность из if-else не подойдет)

03

Какие есть типы задач?

Обучение с учителем (supervised learning)

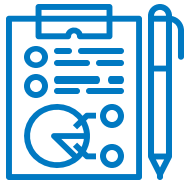
X - множество объектов
Y - множество ответов

$y : X \to Y$ - истинная зависимость

Обучающий датасет - множество наборов из фичей и значений целевой переменной.
Мы обозначим его: c

$$X_{train} \subset X$$

$$X_{train} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ & \dots & \dots & \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \qquad y_{train} = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}$$



Типы признаков (features):

- ❑ Числовые (Numerical)
- ❑ Категориальные (Categorical)
- ❑ Порядковые (Ordinal)

Типы задач:

- ❑ Классификация (Classification)

  $Y = \{0, 1\}, Y = \{1, 2, \dots, n\}, Y = \{0, 1\}^n$

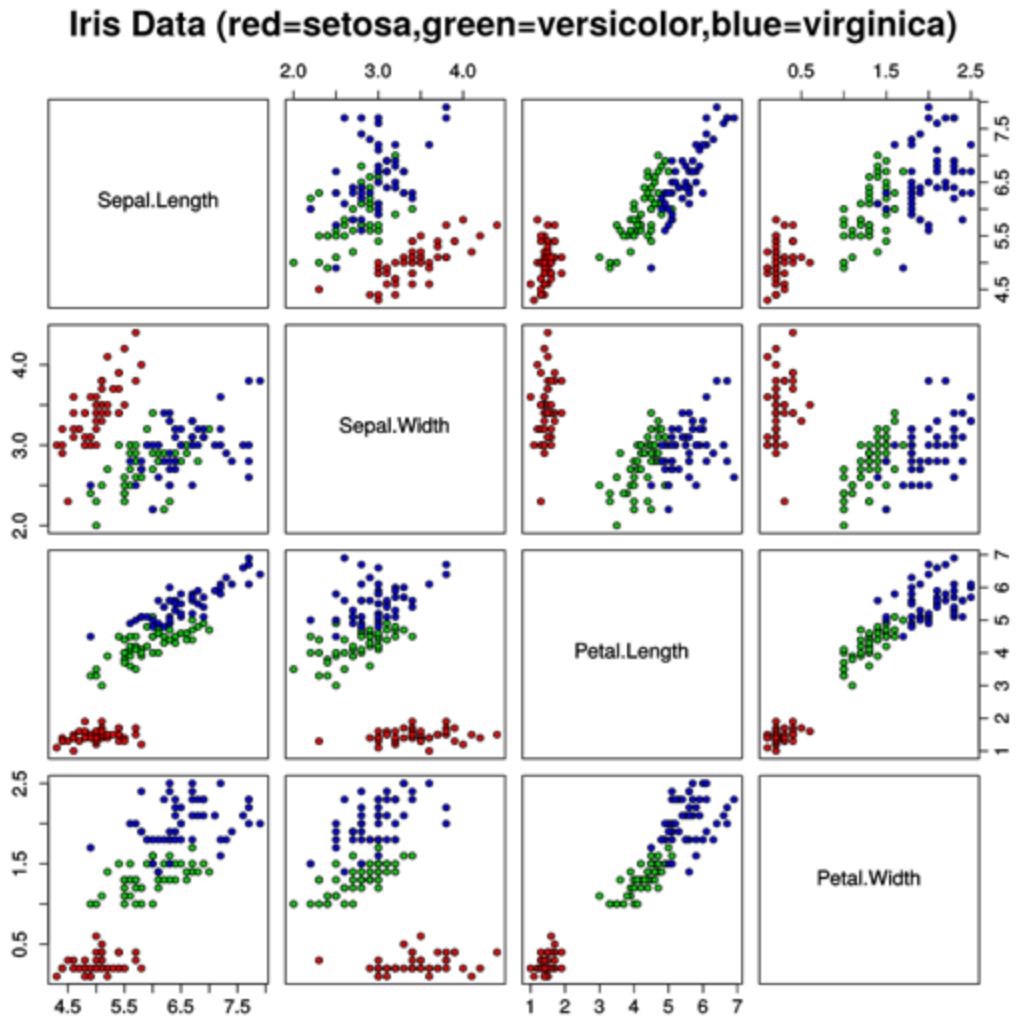
- ❑ Регрессия (Regression)

  $Y = \mathbb{R}$ (числа упорядочены)

- ❑ Ранжирование (Ranking)

  $Y = \{1, 2, \dots, n\}$

Примеры задач (Ирисы Фишера)



Какая это задача?

$$Y = \{1, 2, 3\}$$

Задача классификации.

Какие есть признаки?

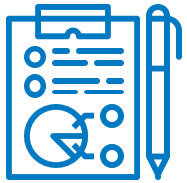
$$X = \mathbb{R}^4$$

Есть только числовые признаки.

Примеры задач (стоимость дома)

Нужно предсказать стоимость дома. Есть обучающий датасет со следующими признаками:

- ✓ Удаленность от метро;
- ✓ Оценка состояния дома (плохое, среднее, хорошее, отличное);
- ✓ Количество комнат;
- ✓ Площадь;
- ✓ Год строительства;
- ✓ Название района, в котором находится дом.



Какая это задача?

$$Y = \mathbb{R}$$

Задача регрессии.

Какие есть признаки?

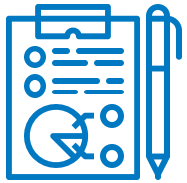
Числовые, порядковые, категориальные.

Примеры задач (поиск страницы в Интернете)

Получив запрос от пользователя нужно найти наиболее полезные документы из некоторой базы.

Что нам известно:

- ✓ Запрос пользователя;
- ✓ Текст документа;
- ✓ Какие ключевые слова есть в каждом документе;
- ✓ Насколько каждый документ популярен.



Какая это задача?

$Y = \{1, 2, \ldots, n\}$ (числа упорядочены)

Задача ранжирования.

Какие есть признаки?

Данные намного сложнее и требуют предобработки.

04
Обучение моделей

Обучение с учителем (supervised learning)

1. Наша задача:
   - найти функцию хорошо приближающую реальную зависимость y(x).
2. Назовем такое решение:
   - $\hat{y} : X \to Y$ (эта функция должна быть вычислима на компьютере).
3. Обычно мы выбираем решение из некоторого параметризованного семейства.
   - $\mathcal{F} = \{\hat{y}_\theta \mid \theta \in \Theta\}, \Theta -$ множество параметров.

Обучение – процесс выбора параметра θ, которому соответствует наиболее подходящее нам решение задачи: $\hat{y}_\theta = (x_1, x_2)$

Пример семейства моделей (функции порога)

1

Задача:

- Определить, можно ли ребенку пройти на аттракцион? Причем мы знаем его рост и возраст.

2

Множество, в котором мы будем искать решения состоит из функций вида:

- $$\hat{y}_{(a,b)}(x_1, x_2) = \begin{cases} 1 & x_1 \geq a, x_2 \geq b \\ 0 & otherwise \end{cases}$$

3

- Параметр в данном случае $\theta = (a, b)$. А множество возможных значений параметра $\Theta = \mathbb{R}$

Обучение с учителем (supervised learning)

1

Функция потерь (loss):

Определим функцию:

$$L(y, \hat{y}(x)),$$

ее значение показывает насколько сильно наше предсказание отличается от реального значения.

2

Пример:

Задача предсказания цены дома из предыдущих примеров.
Возможные функции потерь:

$$L(y_{\text{true}}, \hat{y}(x)) = (y_{\text{true}} - \hat{y}(x))^2$$

(квадратичная функция потерь)

$$L(y_{\text{true}}, \hat{y}(x)) = |y_{\text{true}} - \hat{y}(x)|$$

(абсолютная функция потерь)

3

Эмпирический риск:

Определим эмпирический риск как среднее значение функции потерь на обучающем датасете.
Часто функцию эмпирического риска также называют лоссом.

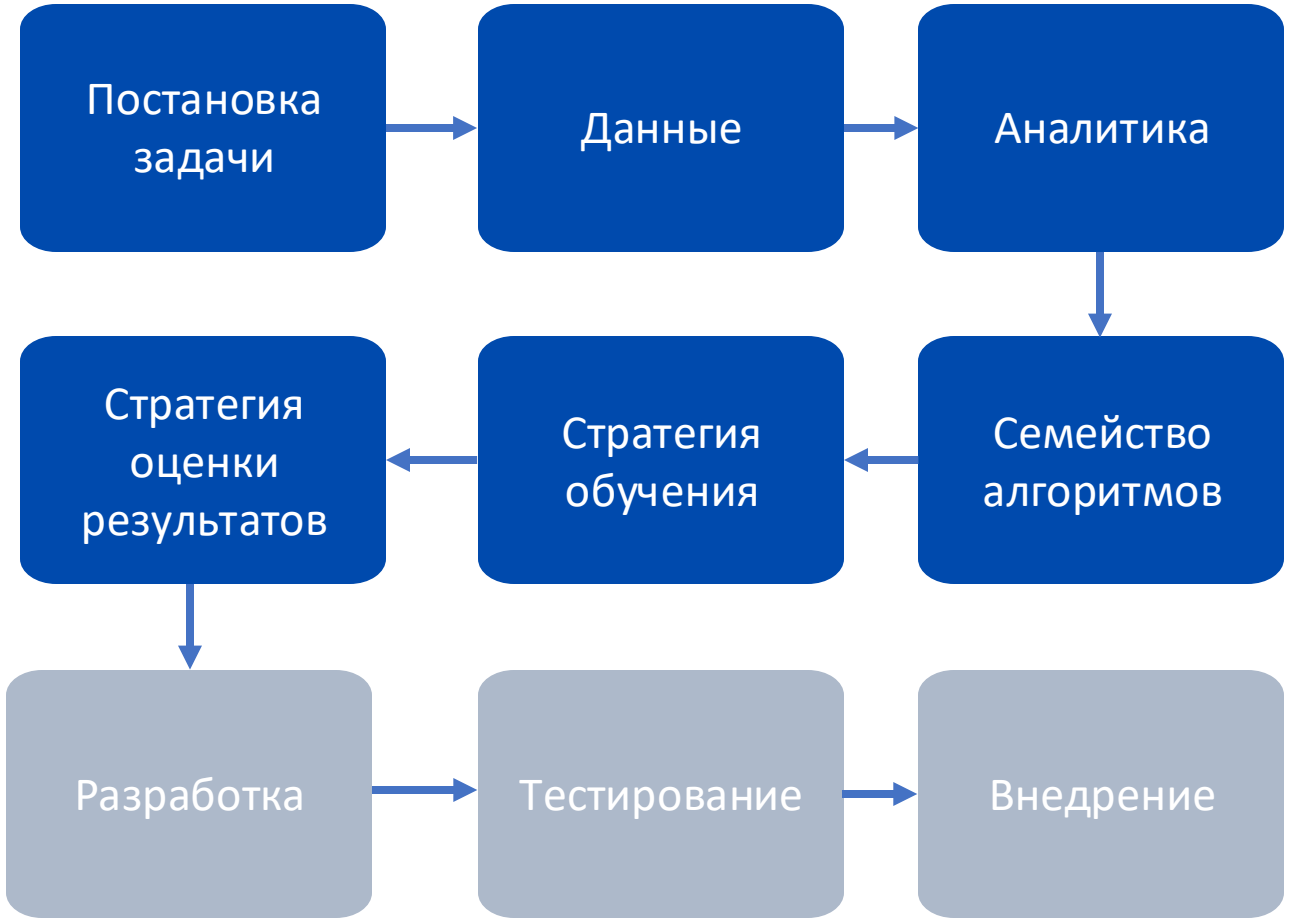
4

Обучение:

$$\theta_{\text{best}} = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{\text{dataset size}} \sum_i L(y^i_{\text{true}}, \hat{y}_\theta(x^i))$$

Это просто математическое определение. Конкретный алгоритм получения лучшего параметра для каждой модели свой.

Резюме



1. У нас есть набор объектов и ответов
2. Из них формируется обучающая выборка
3. Запоминаем примеры
4. Получаем новый объект, сравниваем некоторой функцией
5. Выдаем ответ
6. Радуемся результатам

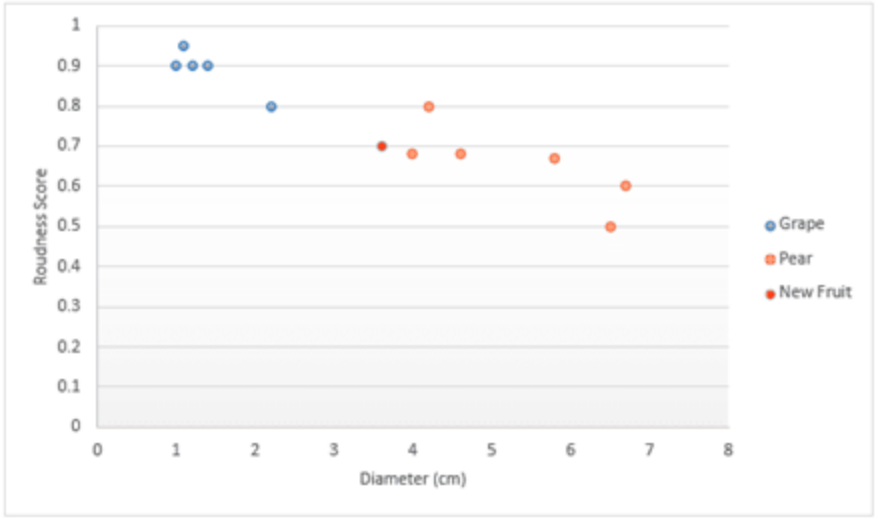
... стоит сейчас задавать вопросы, дальше - сложнее

05
Гипотеза компактности и алгоритм KNN

Гипотеза компактности

- классы образуют компактно локализованные подмножества объектов в пространстве признаков
- похожие объекты находятся "рядом" в пространстве признаков

Пример: классификация фруктов. С какими проблемами мы тут можем столкнуться?



Алгоритм KNN

- Knn - k nearest neighbours
- классифицируем объект исходя из k "похожих" на него

Алгоритм работы

- запоминаем обучающую выборку (да, это всё обучение)
- для каждого нового объекта выводим к какому классу он принадлежит

Алгоритм поиска ответа

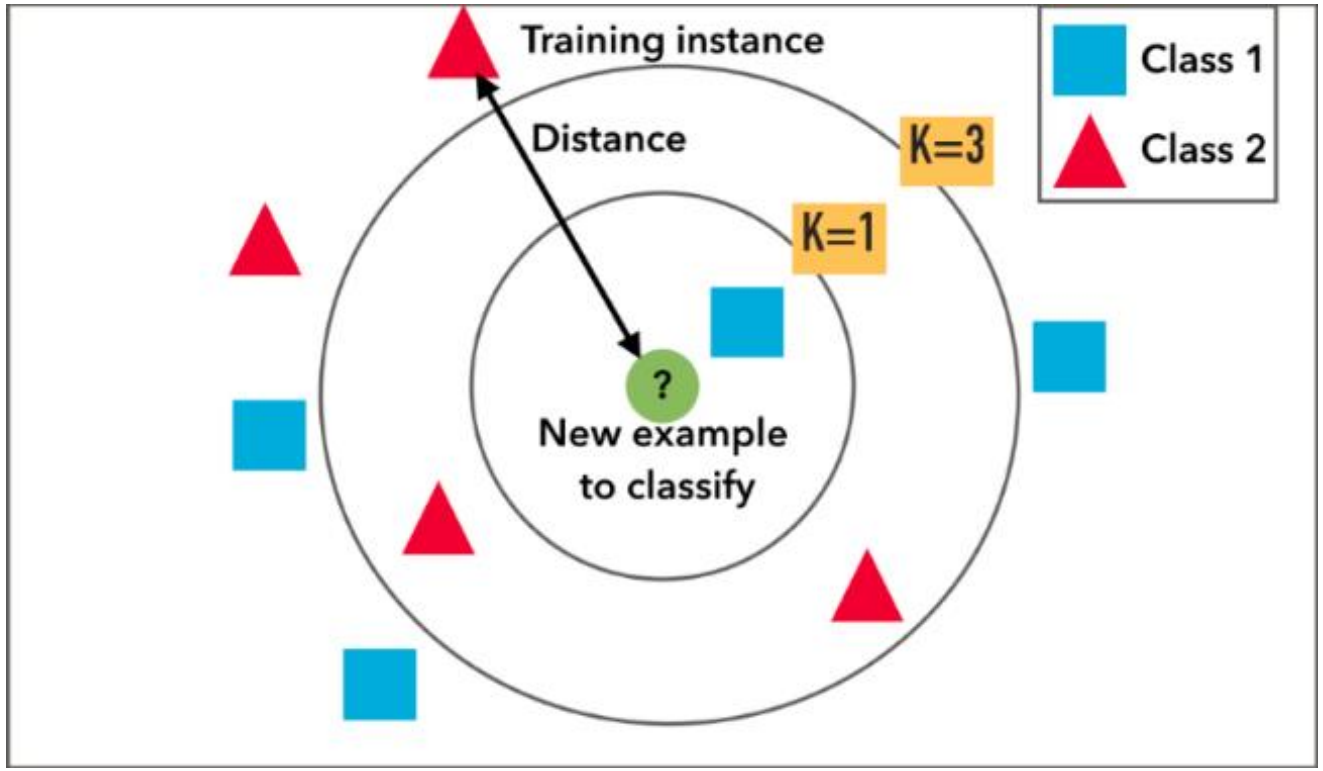
- для нового объекта x считаем расстояние до других объектов

  $\rho(x, x^{(1)}) \leq \rho(x, x^{(2)}) \leq \ldots \leq \rho(x, x^{(l)})$

- сортируем расстояния
- выбираем k ближайших: $x^{(1)}$, $x^{(2)}$, ..., $x^{(l)}$
- выдаём наиболее популярный среди них класс:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} [y_{(i)} = y]$$

Алгоритм KNN



Алгоритм KNN

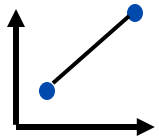
Метрики

- ρ - функция с двумя аргументами
- ρ(x, z) = 0, тогда и только тогда, когда x = z
- ρ(x, z) = ρ(z, x)
- ρ(x, z) ≤ ρ(x, v) + ρ(v, z) - неравенство треугольника

Примеры

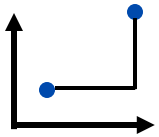
- Евклидова:

$$\rho(x,z) = \sqrt{\sum_{j=1}^{d}(x_j - z_j)^2}$$



- Манхетенская:

$$\rho(x,z) = \sum_{j=1}^{d}|x_j - z_j|$$



- метрика Минковского:

$$\rho(x,z) = \sqrt[p]{\sum_{j=1}^{d}|x_j - z_j|^p}$$

p подбираем под конкретную задачу

06

Наивный Байес

Наивный Байес

Теория

- Теорема Байеса: $P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$

- P(A | B) – вероятность (что A из B истинно)
- P(A) – вероятность (независимая вероятность A)
- P(B | A) – вероятность данного значения признака при данном классе. (что B из A истинно)
- P(B) – вероятность при значении нашего признака. (независимая вероятность B)

$$P(y_i = C_k|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|y_i = C_k)P(y_i = C_k)}{P(\mathbf{x}_i)}$$

$$C^* = \arg\max_k P(y_i = C_k|\mathbf{x_i})$$

$$P(\mathbf{x}_i|y_i = C_k) = \prod_{l=1}^{p} P(x_i^l|y_i = C_k)$$

$$P(y_i = C_k|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|y_i = C_k)P(y_i = C_k)}{P(\mathbf{x}_i)}$$

Наивный Байес

Задача

- определить является ли письмо спамом
- рассчитаем оценку для каждого класса и выберем максимальную по формуле:

$$arg\max[P(Q_k)\prod_{i=1}^{n}P(x_i|Q_k)]$$

$$P(Q_k) = \frac{\text{число документов класса } Q_k}{\text{общее количество документов}}$$

- $P(x_i|Q_k) = \frac{\alpha + N_{ik}}{\alpha M + N_k}$ - вхождение слова x_i в документа класса Q_k
- N_k - количество слов входящих в документ класса Q_k
- M_{ik} - количество слов из обучающей выборки
- N_i - количество вхождений слова x в документ класса Q_k
- α — параметр для сглаживания; мы не можем обучить алгоритм всем словам, и если его не применять, то оценка будет равна 0; $0 < \alpha \leq 1$ (сглаживание Лапласа)

Наивный Байес

Спам

- "Путевки по низкой цене"
- "Акция! Купи шоколадку и получи телефон в подарок"

Не спам

- "Завтра состоится собрание"
- "Купи килограмм яблок и шоколадку"

Требуется определить

- "В магазине гора яблок. Купи семь килограмм и шоколадку"

Наивный Байес

Спам

- Оценка: $\frac{2}{4} \cdot \frac{2}{23} \cdot \frac{2}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \approx 0,000000000587$

Не спам

- Оценка: $\frac{2}{4} \cdot \frac{2}{21} \cdot \frac{2}{21} \cdot \frac{2}{21} \cdot \frac{2}{21} \cdot \frac{1}{21} \cdot \frac{1}{21} \cdot \frac{1}{21} \approx 0,00000000444$

Итог

- Спам < не спам -> письмо не спам!

	Слово	Кол-во вхождений в «Спам»	Кол-во вхождений в «Не спам»	$P(x_i\|\text{Спам})$	$P(x_i\|\text{Не спам})$
Слова из обучающей выборки	Путевки	1	0		
	Низкой	1	0		
	Цене	1	0		
	Акция	1	0		
	Купи	1	1	$\frac{1+1}{14+9}$	$\frac{1+1}{14+7}$
	Шоколадку	1	1	$\frac{1+1}{14+9}$	$\frac{1+1}{14+7}$
	Получи	1	0		
	Телефон	1	0		
	Подарок	1	0		
	Завтра	0	1		
	Состоится	0	1		
	Собрание	0	1		
	Килограмм	0	1	$\frac{1+0}{14+9}$	$\frac{1+1}{14+7}$
	Яблок	0	1	$\frac{1+0}{14+9}$	$\frac{1+1}{14+7}$
	Магазине	0	0	$\frac{1+0}{14+9}$	$\frac{1+0}{14+7}$
	Гора	0	0	$\frac{1+0}{14+9}$	$\frac{1+0}{14+7}$
	Семь	0	0	$\frac{1+0}{14+9}$	$\frac{1+0}{14+7}$

Место для ваших вопросов