Основные понятия в машинном обучении

Данная лекция сделана на основе лекций Воронцова Константина Вячеславовича, которые он читает на кафедре «Интеллектуальные системы» ФУПМ МФТИ.

1 Постановка задачи

1.1 Что задано?

- задано множество объектов $\mathbf{X}$,
- задано множество ответов $\mathbf{Y}$,
- задана некоторая неизвестная функция $\mathbf{f} : \mathbf{X} \to \mathbf{Y}$.

Очевидно, что мы не знаем истиной $\mathbf{f}$. Тогда давайте будем играть в следующую игру, нам дают некоторые $x \in \mathbf{X}$ и дают ответ $y = \mathbf{f}(x)$, а мы хотим найти истинное $\mathbf{f}$.

1.2 Формальная постановка задачи

- дано множество объектов $\{x_1, \cdots x_l\} \subset \mathbf{X}$,
- дано множество ответов $y_i = \mathbf{f}(x_i), i = 1, \cdots, l$,
- найти алгоритм $\mathbf{a} : \mathbf{X} \to \mathbf{Y}$ — приближающий неизвестную функцию $\mathbf{f}$.

Под $\mathbf{a} : \mathbf{X} \to \mathbf{Y}$ понимается некоторая функция, которая каждому объекту из множества $\mathbf{X}$ ставит в соответствие некоторый элемент из $\mathbf{Y}$.

1.3 В чем задача машинного обучения?

- понять как задаются объекты и какими могут быть ответы
- ответить на вопрос в каком смысле $\mathbf{a}$ приближает $\mathbf{f}$
- понять как строить отображение $\mathbf{a}$

На эти 3 вопроса мы и будем пытаться отвечать в течении нашего курса.

2 Задание объектов

Каждый объект выборки задается некоторым набором признаков (features). Что такое признаки? Признаком может быть все что угодно. Например если объект это человек, то в качестве признаков может быть рост, цвет глаз, вес, цвет волос, пол человека и т.д.

2.1 Как задать описание объктов?

$$f_j : X \to D_j, \quad j = 1, \cdots, n, \tag{1}$$

где n это количество признаков для каждого элемента в нашей выборке.

Что такое f_j? f_j это такая функция над объектом x, которая возвращает j-й признак объекта. Например у нас выборка $\mathbf{X}$ это люди, тогда f_j это например функция которая по заданному человеку x возвращает его вес.

Тогда рассмотрим матрицу «объект—признак»:

$$\mathbf{F} = ||f_j(x_i)||_{l \times n} = \begin{bmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \cdots & \cdots & \cdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{bmatrix}. \tag{2}$$

Тоесть матрица $\mathbf{F}$ это матрица которая имеет количество строк равное количеству объектов, а количество столбцов равно количеству признаков.

В нас в курсе у нас всегда изначально будет задана матрица $\mathbf{F}$. Но в общем случаи придумать признаковое описание объектов является достаточно трудной задачей.

3 Задание ответов

Классификация :

- $\mathbf{Y} = \{0, 1\}$ — классификация на 2 класса (если в примере с людьми то м. или ж.),
- $\mathbf{Y} = \{0, 1, \cdots, M\}$ — классификация на M классов (к пример с людьми это цвет волос).

Регрессия :

- $\mathbf{Y} = \mathbb{R}^n$ — в качестве ответом у нас может быть вся числовая ось при $n = 1$.

4 Примеры смысла «a приближает f»

Задача **регрессии**.

Пусть у нас есть множество истинных ответов $\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}$ и множество ответов которые нам дает алгоритм $\mathbf{a}$.

Определим $\delta(\mathbf{a})$ — как ошибку которую допускает алгоритм $\mathbf{a}$ на данном нам множестве $\{x_1, \cdots x_l\} \subset \mathbf{X}$.

$$\delta\mathbf{a} = \sum_{k=1}^{l} ||\mathbf{a}(x_k) - y_k||^2, \tag{3}$$

где под $||\mathbf{a}(x_k) - y_k||$ — подразумевается расстояние от предсказанного ответа до истинного. К примеру, если $Y = \mathbb{R}$, то в качестве расстояния можно взять модуль разности чисел.

Задача **классификации**.

Определим $\delta(\mathbf{a})$ — как ошибку которую допускает алгоритм $\mathbf{a}$ на данном нам множестве $\{x_1, \cdots x_l\} \subset \mathbf{X}$.

$$\delta\mathbf{a} = \sum_{k=1}^{l} [\mathbf{a}(x_k) \neq y_k], \tag{4}$$

где $[\mathbf{a}(x_k) \neq y_k]$ это 1, если условие в скобках истинное и 0 если ложное. Простыми словами $\delta\mathbf{a}$ определили как количество ошибок в классификации.

5 Примеры задач

Задача **регрессии**.

Рассмотрим постановочную задачу регрессии.

5.1 Что дано?

- $\mathbf{X} = \mathbb{R}$ — множество объектов,
- $\mathbf{Y} = \mathbb{R}$ — множество ответов,
- неизвестная функция $\mathbf{f}$ пусть будет просто квадратичная, то есть $y = \mathbf{f}(x) = x^2$

5.2 По каким данным мы восстанавливаем неизвестную функцию?

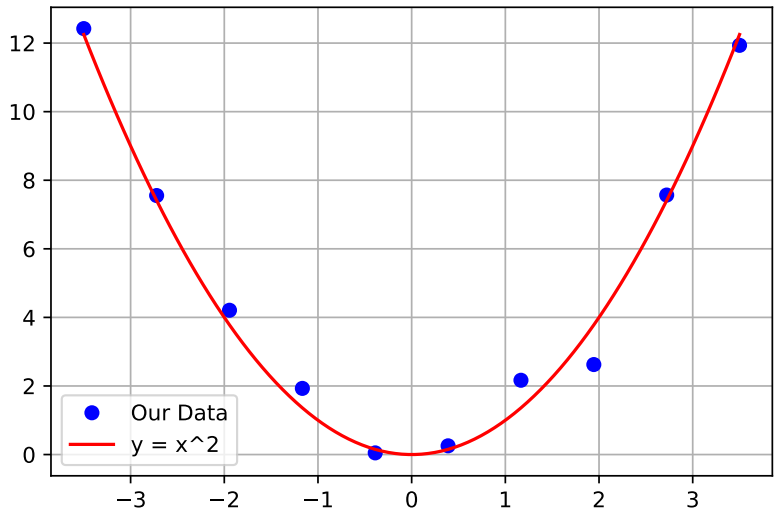

Рис. 1: Данные данные для нахождения $\mathbf{a}$

На рис. 1 показаны данные по которым мы должно построить отображение **a**. По оси абсцисс отложены значения $\{x_1, \cdots, x_{10}\} \subset X$, а по оси ординат отложены значения $\{y_1, \cdots, y_{10}\} \subset Y$. Также на графике построен график функции $y = x^2$. Как видно из графика синие точке не ложатся идеально на красный график. Это все из-за того, что данные не бывают идеальные, об этом мы поговорим чуть позже.

Задача **классификации**.

Рассмотрим постановочную задачу классификации.

5.3 Что дано?

- $\mathbf{X} = \mathbb{R}^2$ — множество объектов,
- $\mathbf{Y} = \{0, 1\}$ — множество ответов,
- неизвестная функция $\mathbf{f}$

5.4 По каким данным мы восстанавливаем неизвестную функцию?

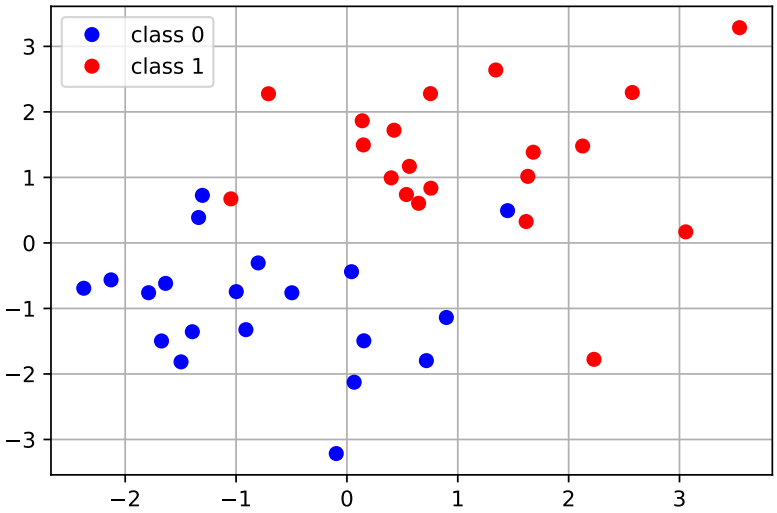

Рис. 2: Данные данные для нахождения **a**

6 Метод k - ближайших соседей

Задача **классификации**.

Данный метод является очень простым. Пусть мы можем измерить расстояние между любыми объектами из множества **X**. Тогда алгоритм заключается в том, чтобы найти тот класс которого больше всего среди $k-$ближайших соседей.

6.1 Алгоритм

k = 9 — сколько соседей будем учитывать
M = 2 — количество классов
$\{x_1, \cdots, x_l\}$ — объекты для которых мы знаем ответы
$\{y_1, \cdots, y_l\}$ — ответы
x — нужно класифицировать
mesure — строим массив расстояний от x до каждого x_i
mesure = mesure.sort — сортируем его по возрастанию (паралельно нужно сортировать и y)
arr — массив счетчик каждого класса среди k ближайших
for i in k
....arr$[y_i]$ = arr$[y_i] + 1$

Список литературы

[1] *Воронцов К. В.* Машинное обучение // Годовой курс кафедры «Интеллектуальные системы» Москва, 2018. `http://www.machinelearning.ru/wiki/index.php?title=Vokov`