Обучение с учителем

Долотов Евгений
dolotov.evgeniy@gmail.com

2017

Аннотация

Рассматривается задача обучения с учителем

Содержание

1 Обучение с учителем

Рассмотрим пример задачи *обучения с учителем* (supervised learning). Пусть у нас есть некоторый набор информации о площади и стоимости домов в опеределенном городе (house_price.csv):

Площадь ($м^2$)	Цена (1000$)
80	190
100	180
105	246
110	214
110	300
⋮	⋮

Для наглядности отобразим эти данные на графике.

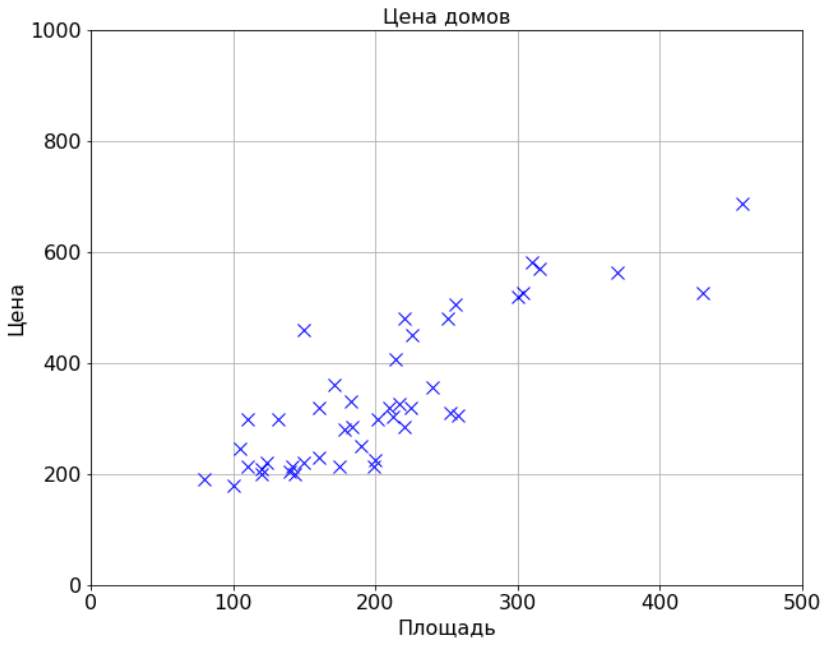

Рис. 1: Зависимость цены дома от его площади

Код для построения графика:

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv(data/house_price.csv)
plt.title(uЦена домов)
plt.axis([0, 500, 0, 1000])
plt.xlabel(uПлощадь)
plt.ylabel(uЦена, a)
plt.grid()
plt.plot(data[area], data[price], bx)
```

Научиться прогнозировать цены на другие жилые дома в этом городе в зависимоти от их площади, опираясь на уже имеющуюся информацию, является типичной задачей обучения с учителем.

1.1 Основные понятия и определения

Пусть задано множество *объектов* X, множество *допустимых ответов* Y, и существует *целевая функция (target function)* $y^*\colon X \to Y$, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1, \ldots, x_l\} \subset X$. Пары «объект-ответ» (x_i, y_i) называются *прецендентами* (training example). Совокупность пар $X^l = (x_i, y_i)_{i=1}^l$ называется *обучающей выборкой* (training set).

Задача обучения с учителем заключается в том, чтобы по выборке X^l восстановить зависимость y^*, то есть построить *решающую функцию* (decision function) $h : X \to Y$, которая приближала бы целевую функцию y^*, причем не только на объектах обучающей выборки, но и на всем множестве X.

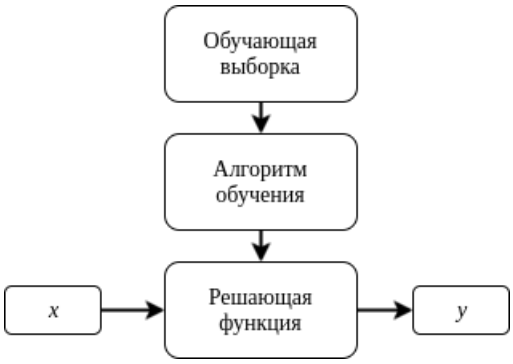

Рис. 2: Принцип решения задачи обучения с учителем

1.1.1 Объекты и признаки

Признак (feature) f объекта x — это результат измерения некоторой характеристики объекта. Формально признаком называется отображение $f : X \to D_f$, где D_f — множество допустимых значений признака. В частности, любой алгоритм $h : X \to Y$ также можно рассматривать как признак.

В зависимости от природы множества D_f признаки делятся на несколько типов:

- Если $D_f = \{0, 1\}$, то f — бинарный признак;
- Если D_f — конечное множество, то f — номинальный признак;
- Если D_f — конечное упорядоченное множество, то f — порядковый признак;
- Если $D_f = R$, то f — количественный признак.

Если все признаки имеют одинаковый тип, $D_{f_1} = \cdots = D_{f_n}$, то исходные данные называются однородными, в противном случае — разнородными.

Пусть имеется набор признаков $f_1, \ldots, f_n$. Вектор $(f_1(x), \ldots, f_n(x))$ называют признаковым описанием объекта $x \in X$. В дальнейшем мы не будем различать объекты из X и их признаковые описания, полагая $X = D_{f_1} \times \ldots \times D_{f_n}$. Совокупность признаковых описаний всех объектов выборки X^l, записанную в виде таблицы размера $l \times n$ называют *матрицей объектов–признаков*:

$$F = \|f_j(x_i)\|_{l \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_l) & \ldots & f_n(x_l) \end{pmatrix}$$

Матрица объектов–признаков является стандартным и наиболее распространённым способом представления исходных данных в прикладных задачах.

1.1.2 Типы задач

В зависимости от природы множества допустимых ответов Y задачи обучения по прецедентам делятся на следующие типы:

- Если $Y = \{1, \ldots, M\}$, то это задача *классификации* (classification) на M непересекающихся классов. В этом случае всё множество объектов X разбивается на классы $K_y = \{x \in X : y(x) = y\}$, и алгоритм $h(x)$ должен давать ответ на вопрос «какому классу принадлежит x?». В некоторых приложениях классы называют *образами* и говорят о задаче *распознавания образов* (pattern recognition).
- Если $Y = \{0, 1\}^M$, то это задача *классификации на M пересекающихся классов*. В простейшем случае эта задача сводится к решению M независимых задач классификации с двумя непересекающимися классами.

- Если $Y = R$, то это задача *восстановления регрессии* (regression estimation).
- Задачи *прогнозирования* (forecasting) являются частными случаями классификации или восстановления регрессии, когда $x \in X$ — описание прошлого поведения объекта x, $y \in Y$ — описание некоторых характеристик его будущего поведения.

1.1.3 Контрольные вопросы

- Определите множество объектов X и множество допустимых ответов Y в рассмотренной выше задаче прогнозирования стоимость дома.
- Какие признаки используются для описания дома? Являются ли исходные данные однородными? Придумайте свои признаки различных типов, которыми можно бы было описать дом.
- К какому типу задач обучения с учителем относится задача определения стоимости дома?
- Подумайте, можно ли рассматривать эту задачу, как задачу классификации?
- Представьте себя на месте человека, обучающего риэлторов. Какими принципами Вы бы посоветовали им воспользоваться при оценке стоимости домов?

1.2 Метод K-ближайших соседей (K-nearest neighbours)