МАШИННОЕ ОБУЧЕНИЕ

Саша Ершова

January 9, 2019

ЗОШ-2019

ПОСТАНОВКА ЗАДАЧИ

Есть выборка объектов. Про каждый объект собраны данные.
Наша задача – выявить такие закономерности в этих данных,
которые можно экстраполировать на генеральную совокупность.

ПРИЗНАКОВОЕ ОПИСАНИЕ ОБЪЕКТА

Допустим, наши объекты — это квартиры:

	Комнаты	Площадь	До метро	Район	Лифт	Цена
1	2	60 кв.м.	3 км	Хамовники	Есть	14 млн
2	3	100 кв.м.	3 км	Бибирево	Есть	20 млн
3	1	35 кв.м.	0.5 км	Внуково	Нет	10 млн

Каждый объект описан вектором; каждая колонка — измерение в пространстве признаков.

И ЧТО С ЭТИМ ДЕЛАТЬ?

	Комнаты	Площадь	До метро	Район	Лифт	Цена
1	2	60 кв.м.	3 км	Хамовники	Есть	14 млн
2	3	100 кв.м.	3 км	Бибирево	Есть	20 млн
3	1	35 кв.м.	0.5 км	Внуково	Нет	10 млн

Выбираем один таргетный признак и пытаемся его предсказать, основываясь на остальных.

И ЧТО С ЭТИМ ДЕЛАТЬ?

	Комнаты	Площадь	До метро	Район	Лифт	Цена
1	2	60	3	5	1	14
2	3	100	3	8	1	20
3	1	35	0.5	35	0	10

Поскольку мы работаем с математическими алгоритмами, всё, что подаётся на вход, так или иначе превратится в числа.

Какие бывают признаки?

Бинарные $D_f = \{0, 1\}$

Номинальные D_f — конечное множество

Порядковые D_f — конечное упорядоченное множество

Количественные D_f — множество действительных чисел

Какие признаки из таблички к какой категории относятся?

БИНАРИЗАЦИЯ ПРИЗНАКОВ

Идея: мы не хотим, чтобы наш алгоритм относился к номинальным признакам как к упорядоченным.

	Цвет глаз
1	Карие
2	Зелёные
3	Голубые

	Цвет глаз
1	1
2	3
3	2

Что использовать?

sklearn.preprocessing.LabelEncoder,
sklearn.preprocessing.OrdinalEncoder

БИНАРИЗАЦИЯ ПРИЗНАКОВ

Идея: мы не хотим, чтобы наш алгоритм относился к номинальным признакам как к упорядоченным, поэтому мы можем перейти от номинального признака к набору бинарных.

	Цвет глаз
1	Карие
2	Зелёные
3	Голубые

	Карие	Голубые	Зелёные
1	1	0	0
2	0	0	1
3	0	1	0

Что использовать?

pandas.get_dummies, sklearn.preprocessing.OneHotEncoder

НОРМАЛИЗАЦИЯ ПРИЗНАКОВ

Иногда если разброс значений количественного признака от 0.5 до 50000000, алгоритм сработает хуже, чем на разбросе от 0 до 1. Поэтому для количественных признаков мы часто хотим перейти к области значений [0, 1]. Для этого от каждого значения признака переходим к его z-score по формуле:

$$z = (x - u)/s$$

Где x — изначальное значение, u — среднее арифметическое выборки, s — стандартное отклонение.

Что использовать?

sklearn.preprocessing.StandardScaler

Обучение с учителем

ЗАДАЧА КЛАССИФИКАЦИИ

Таргетный признак — бинарный или номинальный.

Формальная постановка задачи: есть конечное количество классов, и нужно научиться предсказывать класс для объекта на основе признаков.

Виды классификации

- Бинарная
- Многоклассовая
- Пересекающиеся классы
- Нечёткие классы

Логистическая регрессия

- Изначально решает задачу бинарной классификации
- Алгоритм ищет граничную функцию, которая описывает такую плоскость, которая линейно делит признаковое пространство на две части
- Может быть расширена на многоклассовую задачу
- Плохо работает в случаях, когда пространство признаков нельзя поделить линейно

Что использовать?

sklearn.linear_model.LogisticRegression

Дерево решений

Алгоритм:

1. Выбираем случайный признак из признакового пространства
2. Выбираем такое значение этого признака, которое максимально хорошо поделит наши объекты на два класса
3. Выбираем следующий признак из оставшихся
4. ...
5. Profit!

Что использовать?

sklearn.tree.DecisionTreeClassifier

k-NN классификатор

Алгоритм:

1. Находим k объектов из обучающей выборки, которые ближе всего к нужному.
2. Смотрим, к какому классу относится большинство объектов
3. ...
4. Profit!

Что использовать?

sklearn.tree.DecisionTreeClassifier

Accuracy

accuracy = кол-во правильных ответов / кол-во всех ответов

Предсказание	На самом деле
1	1
2	1
3	3
0	5
2	2
2	3
3	3

Посчитайте accuracy для такой таблички.

F-score

True positive — на самом деле 1, предсказали 1.
False positive — на самом деле 0, предсказали 1.
True negative — на самом деле 0, предсказали 0.
False negative — на самом деле 1, предсказали 0.

F-score

precision = TP / (TP + FP)

Мы присвоили класс 1 какому-то количеству объектов. Какая часть этих объектов действительно относится к 1?

recall = TP / (TP + FN)

В датасете было какое-то количество объектов класса 1, какую их долю мы нашли?

F-score

$$F_1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

F-score

Предсказание: 1, 0, 0, 1, 0, 1, 0, 0, 0, 0
На самом деле: 0, 0, 0, 1, 0, 0, 0, 0, 1, 0

Посчитайте F-score для такого результата.

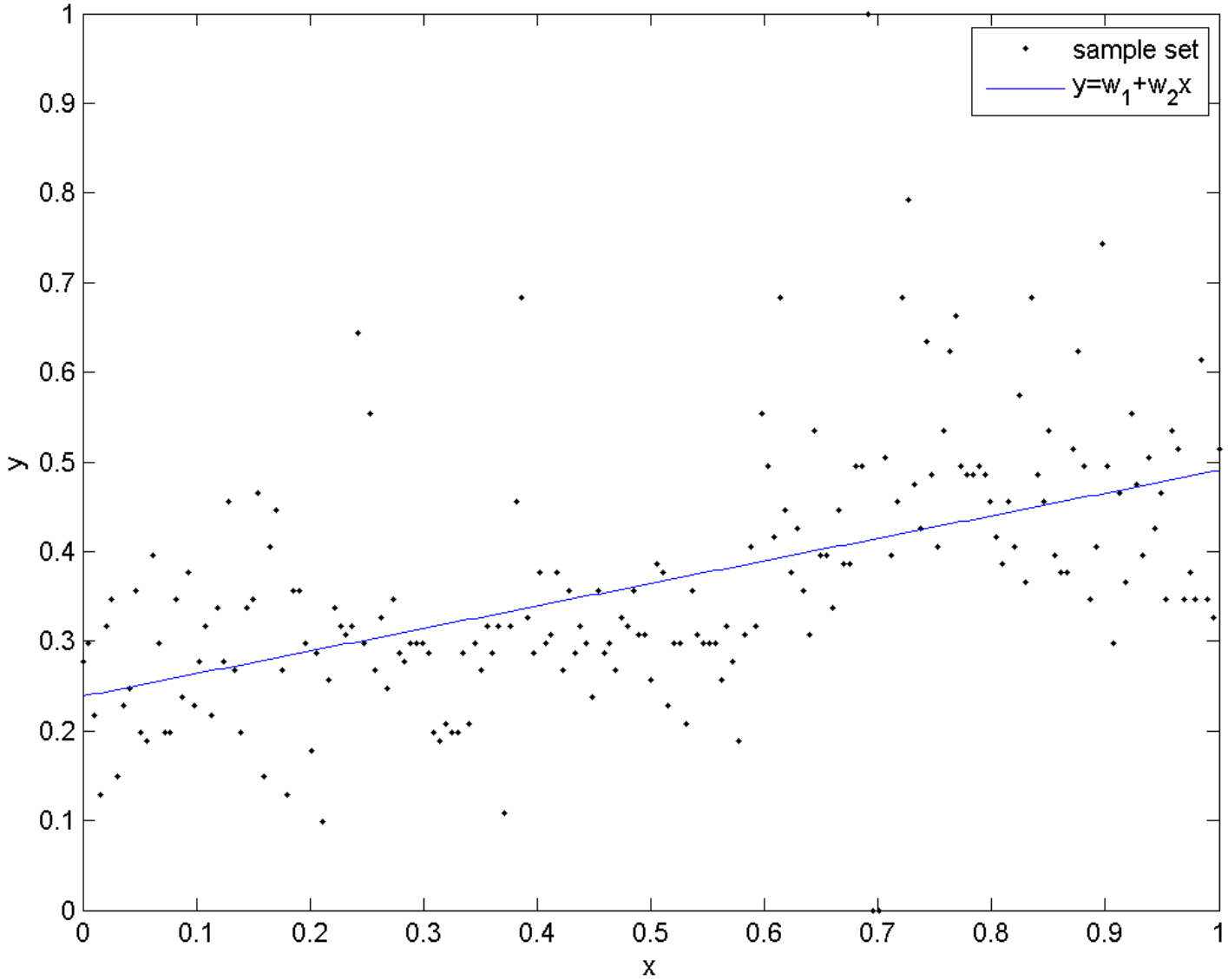
ЗАДАЧА РЕГРЕССИИ

Таргетный признак — количественный, иногда порядковый.

Формальная постановка задачи: по признаковому описанию определить вещественное значение таргетного признака.

Примеры: предсказание оценки фильма или книги, предсказание стоимости квартиры.

Линейная регрессия



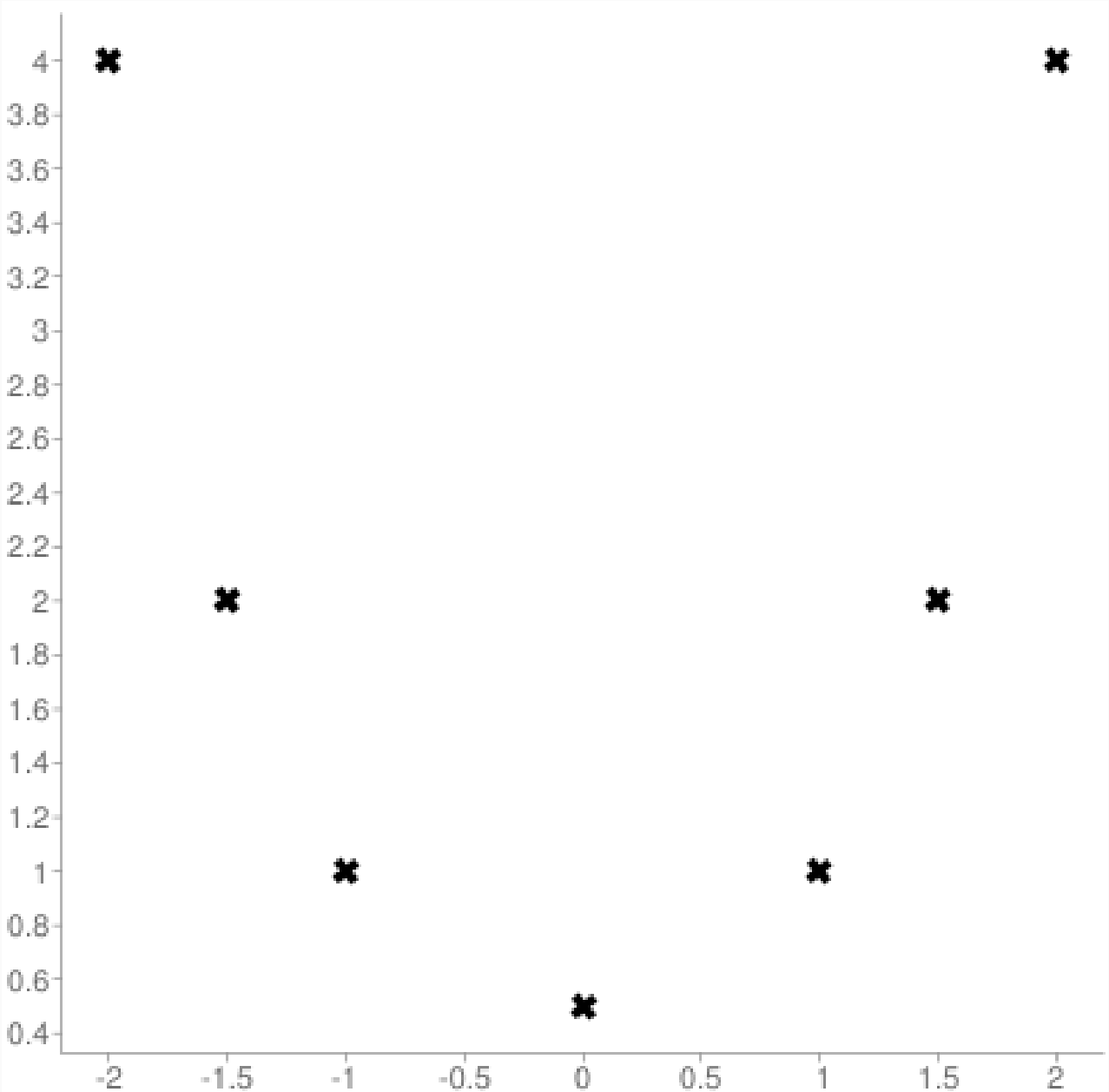
Линейная регрессия

Пусть $x_1, x_2, x_3, x_4, ..., x_n$ — значения признаков в описании объекта, а $w_0, w_1, w_2, w_3, ..., w_n$ — так называемые веса признаков. Тогда мы предполагаем, что значение таргетного признака можно предсказать линейной функцией:

$$y = w_0 + w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n$$

Что использовать?
sklearn.linear_model.LinearRegression

Ограничения линейной регрессии



МЕТРИКИ РЕГРЕССИИ

MSE

Для каждого y из предсказания находим евклидово расстояние до правильного ответа; средний квадрат расстояния — и есть наша метрика.

Что использовать?
sklearn.metrics.mean_squared_error

Спасибо за внимание!