Доклад

Линейные модели

Шалыгин Г. Э.

Российский университет дружбы народов, Москва, Россия

Информация

Докладчик

- Шалыгин Георгий Эдуардович
- студент НФИ-02-20
- Российский университет дружбы народов

Вводная часть

Актуальность

- Предсказание значений признаков, нахождения зависимостей в данных.
- Регрессия, классификация.
- Простота, высокий уровень интерпретируемости, множество алгоритмов построения (обучения)

Цели и задачи

- Изучить линейные модели.
- Задачи:
  - Рассмотреть задачи, решаемые линейными моделями.
  - Изучить линейные модели и варианты их применения в задачах.
  - Исследовать методы улучшения линейных моделей.

Задачи

Регрессия

- Предсказание стоимости квартиры.
- (площадь, этаж, число комнат) -> стоимость
- $M : \mathbb{X} \to \mathbb{R}$

Классификация

- Есть набор операций по банковской карте, а вы бы хотели, понять, какие из этих операций сделали мошенники.
- $M : \mathbb{X} \to 1, 2, \ldots, k, k$ – количество классов.

Модели

Регрессия

- По набору $x_1, x_2, ..., x_d$ предсказываем y.
- Модель:

$$\tilde{y} = w_1 x_1 + ... + w_d x_d + w_0$$

  для какого-то набора w_i.
- Предполагаем, что $y_i \approx \tilde{y}$.
- Уравнение гиперплоскости в пространстве размерности $d + 1$.

Двумерная регрессия

Если признак один, то это прямая:

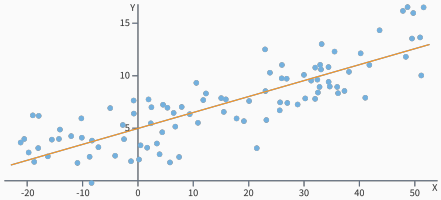

Figure 1: Линейная регрессия

Задача обучения

- Функция потерь: $MSE = \sum_{1}^{n}(y - \tilde{y})^2$.
- Задача минимизации $MSE \to min$.

Возможные решения

- Аналитическое:

$$w = (X^T X)^{-1} X^T y$$

- Приближенное решение:
  - градиентный спуск

$$w_j = w_j - \alpha \frac{d}{dw_j} MSE$$

  - стохастический градиентный спуск (считаем изменение по подвыборке из X).

Классификация

- Есть набор операций по банковской карте, а вы бы хотели, понять, какие из этих операций сделали мошенники.
- $M : \mathbb{X} \to 1, 2, ..., k, k$ – количество классов.
- Модель: $sigm(\tilde{y})$.
- Задача минимизировать кол-во ошибок.

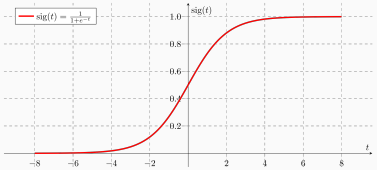

Figure 2: Функция sigm

Дальнейшее изучение моделей

Возможности усовершенствования

- Интерпретация: Цена = 10 x площадь + 1.1 x этаж + 20 x (число комнат).
- Оценка вклада признаков. Тогда признаки нормализуем, коэффициенты покажут их значимость.
- Кроме известных признаков, можно сгенерировать новые:

  Цена = 10 x площадь + 1.1 x этаж + 20 x число комнат − 0.2 x этаж^2 + 0.5 x площадь x число комнат + ☒

Регуляризация

- Признаки линейно зависимы (матрица необратима, коэффициенты не показательны, большая погрешность).
- Изменим функция минимизации:

$$min(MSE + \lambda|w|^k)$$

.

- Для L^2 аналитическое решение:

$$w = (X^T X + \lambda I)^{-1} X^T y$$

Вывод

Вывод

В итоге были рассмотрены линейные модели, имеющие простой вид, высокий уровень интерпретируемости, множество алгоритмов построения и применяющиеся в огромном числе задачах, сводимых к классификации и регрессии.