Обзор неизученных методов классификации и регрессии в scikit-learn

Практикум на ЭВМ 317 группы

Евгений Никишин

ВМК МГУ

17.11.2015

Обоснование:

$$P(y|x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

То есть,

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

- Не нуждается в настройке.
- Применяется при фильтрации спама и классификации текстов.

- Сгенерируем случайную подвыборку с повторением размером N из обучающей выборки.
- Построим решающее дерево, классифицирующее примеры данной подвыборки, причём будем использовать признаки из d случайно выбранных.
- Обученные деревья голосуют.

Важные параметры:

- n_estimators
- criterion
- max_features
- max_depth
- min_samples_leaf
- n_jobs

- Предположим, что признаки стандартизированы
- Старт без переменных
- Ищем x_1, которая наиболее скоррелирована с ответами (то есть имеет наименьший угол)
- Двигаемся в ее направлении до тех пор, пока какая-то другая переменная x_2 не станет столь же коррелируемой
- В данной точке начинаем двигаться в направлении, соответствующему одинаковой корреляции ответов с x_1 и x_2, до тех пор, пока не найдется x_3...

Least Angle Regression

Достоинства:

- Простая модель
- Эффективен при $\mathrm{D} >> \mathrm{N}$
- Быстрый
- Отбор признаков
- Легко настраивать

Недостатки:

- Все недостатки, свойственные итерационным методам
- Неустойчив к шуму

Композиция простых классификаторов. Каждый следующий классификатор строится по объектам, которые плохо классифицируются предыдущими классификаторами.

Достоинства:

- Хорошая обобщающая способность
- Простота реализации
- Время построения композиции определяется временем обучения базовых алгоритмов

Недостатки:

- Чувствителен к шуму и выбросам
- Нужны большие выборки
- Неинтерпретируемость

Бонус: Imputer

Главный параметр — strategy

- mean
- median
- most_frequent

Вопросы?