Тема 2. Регрессия.

Линейная регрессия. Метрики качества.

Epam
Big Data & Data Science

Table of contents

Постановка задачи

Задача регрессии — задача восстановления значения некоторого наблюдаемого показателя в зависимости от известных объясняющих факторов.

$$y = f(x) + \epsilon$$

- y — наблюдаемый показатель
- x — вектор значений объясняющих факторов
- ϵ — случайное слагаемое, представляющее собой совокупное действие остальных факторов, не включенных в модель.

Постановка задачи

Для восстановления значения y строится модель регрессии:

$$\hat{y} = \hat{f}(x)$$

$\hat{f}(x)$ — функция из заранее заданного семейства. Таким семеством могут быть:

- линейные функции: $\hat{f}(x) = \sum_{i=0}^{N} w_i x_i$
- полиномы: $\hat{f}(x) = \sum_{i=0}^{N} w_i x_i^i$
- показательные / степенные функции: $\hat{f}(x) = \sum_{i=0}^{N} e^{w_i x_i}$
- и т.д.

Постановка задачи

В общем виде оценку $\hat{f}(x)$ следует записывать как $\hat{f}(w, x)$, где w - вектор параметров, которые необходимо идентифицировать. Для их идентификации необходимо наличие *обучающей выборки*.

Обучающая выборка — выборка вида $(x_1, y_1) \ldots (x_N, y_N)$, где N — объем выборки (объем массива экспериментальных данных)
y_i, $i = \overline{1, n}$ — наблюдаемые значения показателя
x_i — наблюдаемые значения объясняющих факторов.

МНК - метод наименьших квадратов

Параметры w ищутся как решение оптимизационной задачи

$$\sum_{i=0}^{N}(y_i - \hat{f}(w, x_i))^2 \underset{w}{\longrightarrow} \min$$

Если функцию $\hat{f}(w, x)$ можно представить в виде $\hat{f}(w, x) = Xw$, то решение задачи можно выписать в явном виде:

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

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$$

$$y = (y_1, \ldots, y_N)$$

Метрики качества

Общие метрики:

- Mean squared error: $\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{f}(w, x_i))^2$
- RMSE: $\sqrt{\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{f}(w, x_i))^2}$
- MAE: $\frac{1}{N}\sum_{i=0}^{N} \mid y_i - \hat{f}(w, x_i) \mid$

Статистические метрики качества

- $R^2 = 1 - \frac{\sum_{i=0}^{N}(y_i - \hat{f}(w,x_i))^2}{\sum_{i=0}^{N}((y_i - \overline{y})^2}$
- $R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$, k - число признаков
- F-статистика: $F = \frac{R^2}{1-R^2}(n-2)$
- Анализ регрессионных остатков*

Переобучение

Bias-variance decomposition:

$$E\left[(y-\hat{f}(x))^2\right] = Bias\left[\hat{f}(x)\right]^2 + Var\left[\hat{f}(x)\right] + \varepsilon^2$$

МНК-решение — несмещенная оценка:

$$Bias\left[\hat{f}(x)\right]^2 \ll Var\left[\hat{f}(x)\right]$$

Регуляризация

- $L_1 : \sum_{i=0}^{N}(y_i - \hat{f}(w, x_i))^2 + C \mid w \mid \underset{w}{\longrightarrow} \min$
- $L_1 : \sum_{i=0}^{N}(y_i - \hat{f}(w, x_i))^2 + \frac{C}{2}\|w\|_2 \underset{w}{\longrightarrow} \min$
- Elastic net:
  $\sum_{i=0}^{N}(y_i - \hat{f}(w, x_i))^2 + \lambda \mid w \mid +(1-\lambda)\|w\|_2 \underset{w}{\longrightarrow} \min$

  $\vdots$
- $L_j : \sum_{i=0}^{N}(y_i - \hat{f}(w, x_i))^2 + C\|w\|_j \underset{w}{\longrightarrow} \min, j = \overline{0, \infty}$

Восстановление пропущенных значений

Численные:

- среднее значение
- медиана
- нули
- с помощью модели*

Категориальные:

- самое частое
- отдельной категорией
- с помощью модели*

Представление категориальных переменных

Представление категорий в виде бинарного вектора

Sample:

47	NWAmes
403	SawyerW
900	Sawyer

One hot encoding:

	NWAmes	SawyerW	Sawyer
47	1	0	0
403	0	1	0
900	0	0	1

Нормализация [1]

- StandardScaler: $\hat{x} = \frac{x - \overline{x}}{\sigma^2}$
- MinMaxScaler: $x \rightarrow [0, 1]$
- Робастная нормализация (квантили, логарифмирование и т.д.)

[0] http://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

Оценка качества

Train-test split

- 70% - train, 30% - test
- Стратификация

Cross-validation

- K-частей
- Стратификация
- Перемешивание
- Качество — средняя оценка

Оценка качества

Train-test split

- 70% - train, 30% - test
- Стратификация

Cross-validation

- K-частей
- Стратификация
- Перемешивание
- Качество — средняя оценка

NB: Для данных, упорядоченных по времени, разбиение должно быть также по времени

Полезные ссылки

- http://bjlkeng.github.io/posts/a-probabilistic-view-of-regression/ - вероятностная интерпретация модели линейной регрессии
- http://cs229.stanford.edu/notes/cs229-notes1.pdf - обобщенные линейные модели (GLM)L