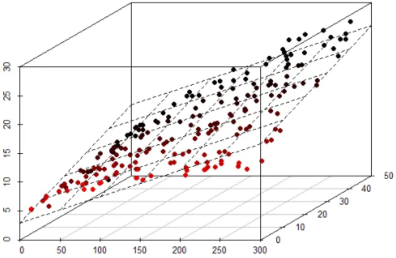
Линейная регрессия и обобщения

Виктор Китов

victorkitov.github.io

Курс поддержан фондом 'Интеллект'



Победитель конкурса VK среди курсов по IT

Содержание

Линейная регрессия

- Линейная регрессия

$$\widehat{y} = x^T \widehat{w} = \sum_{i=1}^{D} \widehat{w}_i x^i$$

$$\widehat{w} = \arg\min_{w} \sum_{n=1}^{N} \left(x_n^T w - y_n \right)^2$$

- Если смещение $\widehat{w}_0$ явно не указано, всегда включают константный признак в x.
- Предположения:

Линейная регрессия

- Линейная регрессия

$$\widehat{y} = x^T \widehat{w} = \sum_{i=1}^{D} \widehat{w}_i x^i$$

$$\widehat{w} = \arg\min_w \sum_{n=1}^{N} \left(x_n^T w - y_n \right)^2$$

- Если смещение $\widehat{w}_0$ явно не указано, всегда включают константный признак в x.
- Предположения:
    - каждый x^i линейно влияет y с коэффициентом $\widehat{w}_i$
    - вклад каждого признака x^i не зависит от значений др. признаков.

Анализ метода

Преимущества:

- интерпретируемость
  - знак коэффициентов=направление влияния x^i
  - модуль коэффициента=сила влияния x^i (при признаках из одной шкалы!)
  - $\widehat{w}$ асимптотически нормальны (см. ссылку), можем тестировать:
    - значимость отличия коэффициентов (или группы коэффициентов) от нуля,
    - гипотезу положительного влияния признака на отклик (положительности коэффициента)
  - есть аналитическое решение
  - быстро и просто строятся прогнозы
  - меньше переобучается, чем сложные модели
    - для больших D может быть оптимальной моделью

Анализ метода

Преимущества:

- интерпретируемость
    - знак коэффициентов=направление влияния x^i
    - модуль коэффициента=сила влияния x^i (при признаках из одной шкалы!)
    - $\widehat{w}$ асимптотически нормальны (см. ссылку), можем тестировать:
        - значимость отличия коэффициентов (или группы коэффициентов) от нуля,
        - гипотезу положительного влияния признака на отклик (положительности коэффициента)
    - есть аналитическое решение
    - быстро и просто строятся прогнозы
    - меньше переобучается, чем сложные модели
        - для больших D может быть оптимальной моделью

Недостатки: модельные предположения слишком простые

- признаки могут влиять нелинейно
- признаки могут иметь взаимозависимое влияние

Решение

Метод наименьших квадратов (МНК, ordinary least squares):

$$L(w) = \sum_{n=1}^{N} \left(x_n^T w - y_n \right)^2 = \|Xw - Y\|_2^2 \to \min_w, \quad X \in \mathbb{R}^{NxD}$$

Решение

Метод наименьших квадратов (МНК, ordinary least squares):

$$L(w) = \sum_{n=1}^{N} \left(x_n^T w - y_n \right)^2 = \|Xw - Y\|_2^2 \to \min_w, \quad X \in \mathbb{R}^{NxD}$$

$$\nabla L(\widehat{w}) = 2 \sum_{n=1}^{N} x_n \left(x_n^T \widehat{w} - y_n \right) = 0$$

$$\left(\sum_{n=1}^{N} x_n x_n^T \right) \widehat{w} = \sum_{n=1}^{N} x_n y_n$$

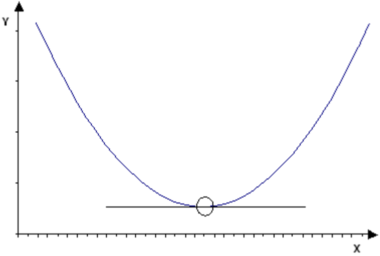
$$X^T X \widehat{w} = Y$$

$$\widehat{w} = (X^T X)^{-1} X^T Y$$

$\widehat{w}_i \propto$ковариации x_n^i и y_n, нормализованной на $Var[x^i]$,$cov[x^i, x^j]$.

Глобальность минимума

- Это глобальный минимум, т.к. оптимизируемый критерий выпуклый.
  - выпуклая ф-ция от линейной выпукла[1], сумма выпуклых - выпукла
  - для выпуклой ф-ции достаточное условие минимума - равенство нулю производной.



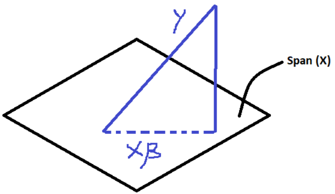
[1] Будет ли суперпозиция произвольных выпуклых ф-ций выпуклой?

Геометрическая интерпретация

- Находится линейная комбинация признаков, чтобы приблизить Y в $\mathbb{R}^N$:

$$L(w) = \sum_{n=1}^{N} \left(x_n^T w - y_n \right)^2 = \|Xw - Y\|_2^2 \to \min_w$$

- Решение - проекция на линейную оболочку признаков в $\mathbb{R}^N$.



Линейно зависимые признаки - проблема

- Решение $\widehat{w} = (X^T X)^{-1} X^T Y$ существует, когда $X^T X$ невырождена.
- Поскольку $rank(X) = rank(X^T X)\ \forall X$, проблема возникает при линейной зависимости признаков.

Линейно зависимые признаки - проблема

- Решение $\widehat{w} = (X^T X)^{-1} X^T Y$ существует, когда $X^T X$ невырождена.
- Поскольку $rank(X) = rank(X^T X)\ \forall X$, проблема возникает при линейной зависимости признаков.
    - пример: константный признак и one-hot закодированные $e_1, e_2, ...e_K$, поскольку $\sum_k e_k \equiv 1$
    - интерпретация: возникает неоднозначность $\widehat{w}$ для зависимых признаков:
        - линейная зависимость: $\exists \alpha :\ x^T \alpha = 0\, \forall x$
        - предположим $\widehat{w}$ - решение $\sum_{n=1}^{N} \left(x_n^T w - y_n\right)^2 \to \min_w$
        - тогда $\widehat{w} + k\alpha$ - тоже решение $\forall k \in \mathbb{R}$:$x^T \widehat{w} \equiv x^T \widehat{w} + k x^T \alpha \equiv x^T (\widehat{w} + k\alpha)$.
- При почти зависимых признаках ($X^T X$ плохо обусловлена, т.е. $\lambda_{max}/\lambda_{min}$ велико):
    - $\widehat{w}$ неустойчиво и принимает большие по модулю значения.

Линейно зависимые признаки - решение

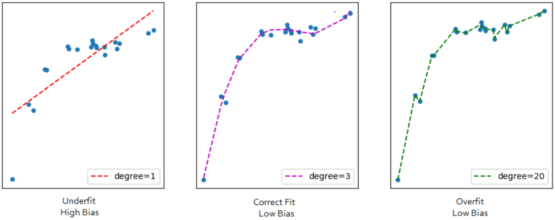
- Проблема может быть решена:
  - отбором признаков (feature selection)
  - снижением размерности (dimensionality reduction)
  - накладыванием доп. условий на решение (регуляризация)
    - $\|w\|$ должна быть мала
    - некоторые w_i должны быть неотрицательные
    - ...

Нелинейные зависимости в линейной регрессии

Перейдем от $x \in \mathbb{R}^D$ к его нелинейному преобразованию $\in \mathbb{R}^M$:

$$x \to [\phi_1(x), \phi_2(x), \dots \phi_M(x)]$$

$$\widehat{y}(x) = \phi(x)^T \widehat{w} = \sum_{m=1}^{M} \widehat{w}_m \phi_m(x)$$



Лин. регрессия с полиномиальным преобразованием:
$x \to [x, x^2, x^3, \dots x^{degree}]$

Анализ

$\widehat{y}(x)$ уже нелинейно зависит от x. При этом преимущества лин. регрессии сохраняются:

- интерпретируемость (для несложных преобразований)
- аналитическое решение
- глобальный минимум потерь

Нелинейная регрессия

- Можно исходные признаки подставлять в нелинейную ф-цию $\widehat{y} = f(x|w)$

$$L(w|X, Y) = \sum_{n=1}^{N} (f(x_n|w) - y_n)^2$$

$$\widehat{w} = \arg\min_{w} L(w|X, Y)$$

- В общем случае не существует аналитического решения $\widehat{w}$.
  - используем численные методы, например SGD.

Пример использования

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error

X_train, X_test, Y_train, Y_test =
    get_demo_regression_data()
model = LinearRegression()     # инициализация модели
model.fit(X_train,Y_train)     # обучение модели
Y_hat = model.predict(X_test) # построение прогнозов
print(f'Средний модуль ошибки (MAE): \
   {mean_absolute_error(Y_test, Y_hat):.2f}')
```

Больше информации. Полный код.

Содержание

Регуляризация

- Для лучшей обобщающей способности важна не только точность, но и простота модели.
- Учтем простоту дополнительным регуляризатором $R(w)$:

$$\sum_{n=1}^{N}\left(x_n^T w - y_n\right)^2 + \lambda R(w) \to \min_w$$

- $\lambda > 0$ - гиперпараметр[2], контролирующий сложность модели.

$$\begin{array}{ll} R(w) = ||w||_1, & \text{Лассо регрессия (Lasso regression)} \\ R(w) = ||w||_2^2 & \text{Гребневая регрессия (Ridge regression)} \end{array}$$

- На практике смещение часто не регуляризуют, чтобы не приводить к смещению прогнозов к нулю.

[2] Как он влияет на сложность модели?

Пример использования гребневой регрессии

```
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_absolute_error

X_train, X_test, Y_train, Y_test =
    get_demo_regression_data()
model = Ridge(alpha=1)     # инициализация модели
model.fit(X_train,Y_train)      # обучение модели
Y_hat = model.predict(X_test)   # построение прогнозов
print(f'Средний модуль ошибки (MAE): \
   {mean_absolute_error(Y_test, Y_hat):.2f}')
```

- α - вес при регуляризаторе (а не при ф-ции потерь).
- Больше информации. Полный код.

Пример использования LASSO регрессии

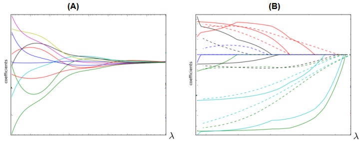
```
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_absolute_error

X_train, X_test, Y_train, Y_test =
    get_demo_regression_data()
model = Lasso(alpha=1)      # инициализация модели
model.fit(X_train,Y_train)        # обучение модели
Y_hat = model.predict(X_test)     # построение прогнозов
print(f'Средний модуль ошибки (MAE): \
   {mean_absolute_error(Y_test, Y_hat):.2f}')
```

- α - вес при регуляризаторе (а не при ф-ции потерь).
- Больше информации. Полный код.

Зависимость $\widehat{w}$ от λ

- Зависимость $\widehat{w}$ от λ для гребневой (A) и лассо (B) регрессии:



- Лассо регрессия может использоваться для автоматического отбора признаков.
- λ находят по экспоненциальной сетке $[10^{-6}, 10^{-5}, ...10^{5}, 10^{6}]$.
    - потом уточняют
- Всегда рекомендуется включать регуляризацию:
    - плавный контроль сложности модели
    - решение однозначно даже для линейно зависимых признаков
        - из набора решений выбирается с наименьшим $\|w\|$.

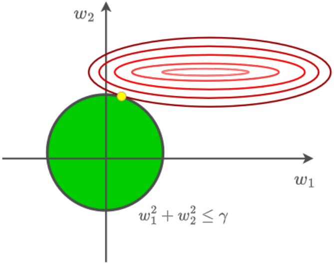
Разное поведение L1 и L2 регуляризации

Разное поведение L1 и L2 регуляризации объясняется эквивалентностью следующих оптимизационных задач:

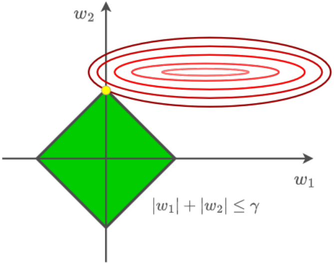
$$L(w) + \lambda R(w) \to \min_w \iff \begin{cases} L(w) \to \min_w \\ R(w) \le \gamma \end{cases}$$

где $\gamma = \gamma(\lambda)$ и доказывается из условий Каруша-Куна-Таккера.

Оптимизация при L2 регуляризации



Оптимизация при L1 регуляризации



ElasticNet

- ElasticNet - линейная комбинация L_1 и L_2 регуляризации:

$$R(w) = \alpha||w||_1 + (1-\alpha)||w||_2^2$$

$$\alpha \in [0,1] - \text{гиперпараметр.}$$

- Если два признака x^i и x^j равны:
  - Гребневая регрессия выберет оба с равным весом
    - правильно, т.к. нет априорных предпочтений
  - Лассо регрессия выберет один из них (в общем случае)
    - зато отберет лишние признаки
- ElasticNet обладает обоими преимуществами.

Аналитическое решение для гребневой регрессии

Критерий гребневой регрессии

$$\sum_{n=1}^{N}\left(x_n^T w - y_n\right)^2 + \lambda w^T w \to \min_w$$

Условие стационарности (равенство нулю производной):

Аналитическое решение для гребневой регрессии

Критерий гребневой регрессии

$$\sum_{n=1}^{N}\left(x_n^T w - y_n\right)^2 + \lambda w^T w \to \min_w$$

Условие стационарности (равенство нулю производной):

$$2\sum_{n=1}^{N} x_n\left(x_n^T \widehat{w} - y_n\right) + 2\lambda\widehat{w} = 0$$

$$2X^T(X\widehat{w} - Y) + 2\lambda\widehat{w} = 0$$

$$\left(X^T X + \lambda I\right)\widehat{w} = X^T Y$$

поэтому

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

$X^T X + \lambda I$ всегда невырождена как сумма $X^T X \succeq 0$ и $\lambda I \succ 0$.

Зашумление признаков

- Приём регуляризации: зашумление признаков во время обучения модели с шумом $\delta \in \mathbb{R}^D$:

$$x \to x + \delta$$

- Шум генерируется свой на каждом шаге оптимизации и удовлетворяет
$$\mathbb{E}\delta = 0, \quad \mathbb{E}\delta = \lambda I$$
- Во время применения модели признаки не зашумляются.
- Препятствуем модели сильно полагаться на отдельный признак и учитывать его с большой силой.
- Это общий приём для любой модели.
- В случае линейной регрессии он эквивалентен L2 регуляризации.

Эквивалентность зашумления и L2 регуляризации

Усреднённый MSE по всевозможным реализациям шума:

Эквивалентность зашумления и L2 регуляризации

Усреднённый MSE по всевозможным реализациям шума:

$$
\begin{aligned}
L(w) &= \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\right\} = \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-(x_i+\delta_i)^T w)^2\right\} \\
&= \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}((y_i-x_i^T w)-\delta_i^T w)^2\right\} \\
&= \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-x_i^T w)^2-2\delta_i^T w(y_i-x_i^T w)+w^T\delta_i\delta_i^T w\right\} \\
&= \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-x_i^T w)^2\right\}-2\mathbb{E}\left\{\delta_i^T w(y_i-x_i^T w)\right\}+\mathbb{E}\left\{w^T\delta_i\delta_i^T w\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i-x_i^T w)^2+\lambda\|w\|_2^2,
\end{aligned}
$$

Учет разных признаков с разной силой

- При масштабированию признаков прогнозы лин. регрессии

Учет разных признаков с разной силой

- При масштабировании признаков прогнозы лин. регрессии не изменятся:

$$\widehat{y} = \widehat{w}_1 x^1 + \widehat{w}_2 x^2 + \ldots \xrightarrow{x^1 \to x^1/\alpha} (\alpha \widehat{w}_1)\left(\frac{x^1}{\alpha}\right) + \widehat{w}_2 x^2 + \ldots$$

- А с регуляризацией изменятся:

$$\sum_{n=1}^{N} \left(x_n^T w - y_n\right)^2 + \lambda R(w) \to \min_w$$

- После изменения масштаба признаков, они будут вносить другой вклад в прогноз.
    - для большего учета признака как нужно изменить его масштаб?

Содержание

Обобщение функции потерь[3]

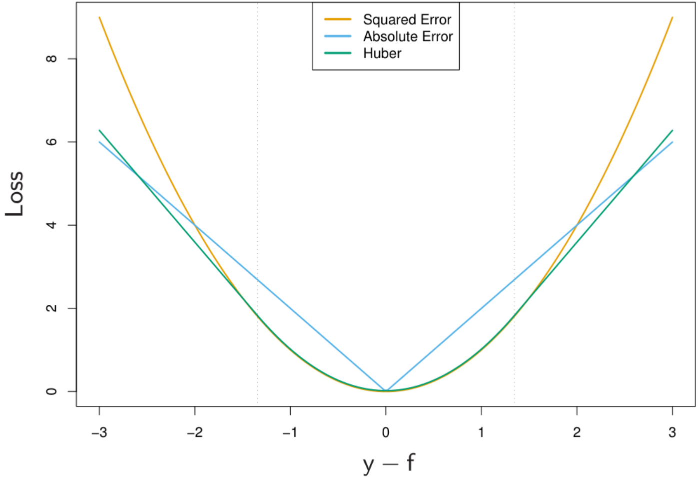
- Обобщим квадратичные потери на произвольные:

$$\sum_{n=1}^{N}\left(x^T w - y_n\right)^2 \to \min_w \qquad \Longrightarrow \qquad \sum_{n=1}^{N} \mathcal{L}(x_n^T w - y_n) \to \min_w$$

ФУНКЦИЯ ПОТЕРЬ	НАЗВАНИЕ	СВОЙСТВА
$\mathcal{L}(\varepsilon) = \varepsilon^2$	квадратичная	дифференцируемая
$\mathcal{L}(\varepsilon) = \lvert\varepsilon\rvert$	абсолютная	устойчивая к выбросам
$\mathcal{L}(\varepsilon) = \begin{cases} \frac{1}{2}\varepsilon^2, & \lvert\varepsilon\rvert \le \delta \\ \delta\left(\lvert\varepsilon\rvert - \frac{1}{2}\delta\right) & \lvert\varepsilon\rvert > \delta \end{cases}$	Хубера	оба свойства

[3] Чему равен константный прогноз, минимизирующий квадратичные и абсолютные ошибки?

Визуализация функций потерь



Оптимальный прогноз для квадратичной ошибки

Константный прогноз $\widehat{y} \in \mathbb{R}$ при квадратичной ф-ции потерь:

$$L(\widehat{y}) = \mathbb{E}\left\{(\widehat{y} - y)^2\right\} \to \min_{\widehat{y} \in \mathbb{R}}$$

Оптимальный прогноз для квадратичной ошибки

Константный прогноз $\widehat{y} \in \mathbb{R}$ при квадратичной ф-ции потерь:

$$L(\widehat{y}) = \mathbb{E}\left\{(\widehat{y} - y)^2\right\} \to \min_{\widehat{y} \in \mathbb{R}}$$

$$\frac{\partial L(\widehat{y})}{\partial \widehat{y}} = \mathbb{E}\left\{2(\widehat{y} - y)\right\} = 2\widehat{y} - 2\mathbb{E}y = 0$$

$$\widehat{y} = \mathbb{E}y$$

Оптимальный прогноз для абсолютной ошибки

Константный прогноз $\widehat{y} \in \mathbb{R}$ при абсолютной ф-ции потерь:

$$L(\widehat{y}) = \mathbb{E}\left\{|\widehat{y} - y|\right\} = \int |\widehat{y} - y|\, p(y) dy =$$

$$= \int (\widehat{y} - y)\mathbb{I}[\widehat{y} \geq y] p(y) dy + \int (y - \widehat{y})\mathbb{I}[\widehat{y} < y] p(y) dy \to \min_{\widehat{y} \in \mathbb{R}}$$

Оптимальный прогноз для абсолютной ошибки

Константный прогноз $\widehat{y} \in \mathbb{R}$ при абсолютной ф-ции потерь:

$$L(\widehat{y}) = \mathbb{E}\left\{|\widehat{y} - y|\right\} = \int |\widehat{y} - y|\, p(y)dy =$$

$$= \int (\widehat{y} - y)\mathbb{I}[\widehat{y} \geq y]p(y)dy + \int (y - \widehat{y})\mathbb{I}[\widehat{y} < y]p(y)dy \to \min_{\widehat{y} \in \mathbb{R}}$$

$$\frac{\partial L(\widehat{y})}{\partial \widehat{y}} = \int \mathbb{I}[\widehat{y} \geq y]p(y)dy - \int \mathbb{I}[\widehat{y} < y]p(y)dy = 0$$

$$\frac{\partial L(\widehat{y})}{\partial \widehat{y}} = \int_{y \leq \widehat{y}} p(y)dx - \int_{y > \widehat{y}} p(y)dy = 0$$

$$\widehat{y} = \text{median}[y]$$

Влияние функции потерь на результат

- Следовательно, для фиксированного x оптимальный функциональный прогноз будет:

$$\arg\min_{\widehat{y}(x)} \mathbb{E}\left\{ (\widehat{y}(x) - y)^2 \middle| x \right\} = \mathbb{E}[y|x]$$

$$\arg\min_{\widehat{y}(x)} \mathbb{E}\left\{ |\widehat{y}(x) - y| \, | \, x \right\} = \text{median}[y|x]$$

- При фиксированных обучающей выборке и модели результат будет получаться разный для различных ф-ций потерь!

Содержание

Взвешенный учет наблюдений[4]

- Взвешенный учет наблюдений

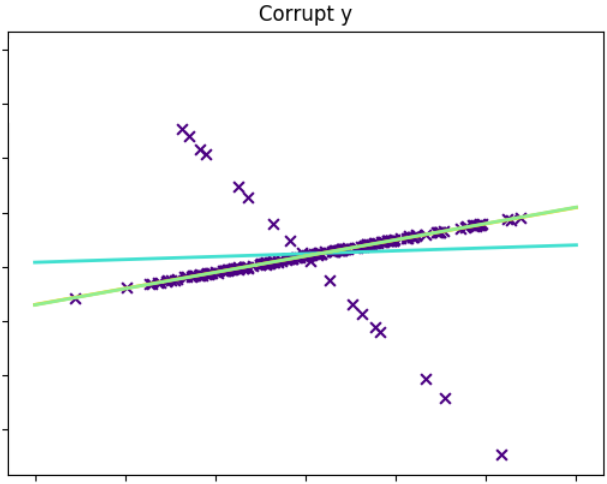
$$\sum_{n=1}^{N} \alpha_n (x_n^T w - y_n)^2 \to \min_{w \in \mathbb{R}^D}$$

$$\alpha_1 \geq 0, \dots \alpha_N \geq 0$$

- Неравномерные веса могут быть обусловлены:
    - разному доверию различным фрагментам обучающей выборки
    - желанием снизить влияние объектов-выбросов
    - желанием сделать сделать сбалансированную выборку
        - Например, при голосовании женщины голосовали чаще мужчин. Но хотим универсальную модель для мужчин и женщин.

[4] Выведите решение для взвешенной линейной регрессии.

Проблема выбросов



Робастная регрессия

- Инициализировать $\alpha_1 = ... = \alpha_N = 1/N$
- Повторять до сходимости:
  - оценить регрессию $\widehat{y}(x)$ используя (x_i, y_i) с весами α_i.
  - для каждого $i = 1, 2, ...N$:
    - переоценить $\varepsilon_i = \widehat{y}(x_i) - y_i$
    - пересчитать веса $\alpha_i = K\left(|\varepsilon_i|\right)$
  - нормализовать веса $\alpha_i = \frac{\alpha_i}{\sum_{n=1}^{N} \alpha_n}$

Комментарии:

- $K(\cdot)$ - некоторая убывающая функция.
- Веса объектов-выбросов убывают, получаем устойчивое к выбросам решение.
- Алгоритм обобщается на любой метод, допускающий взвешенный учет наблюдений.

Orthogonal matching pursuit: задача

Метод Orthogonal Matching Pursuit решает задачу:

$$\begin{cases} \|Xw - Y\|_2^2 \to \min_w \\ \|w\|_0 \le K \end{cases}$$

или эквивалентную (для $\varepsilon = f(K)$ для некоторой $\downarrow f(\cdot)$):

$$\begin{cases} \|w\|_0 \to \min_w \\ \|Xw - Y\|_2^2 \le \varepsilon \end{cases}$$

- $\|w\|_0 = \#[\text{число ненулевых весов}]$

Orthogonal matching pursuit: задача

Метод Orthogonal Matching Pursuit решает задачу:

$$\begin{cases} \|Xw - Y\|_2^2 \to \min_w \\ \|w\|_0 \le K \end{cases}$$

или эквивалентную (для $\varepsilon = f(K)$ для некоторой $\downarrow f(\cdot)$):

$$\begin{cases} \|w\|_0 \to \min_w \\ \|Xw - Y\|_2^2 \le \varepsilon \end{cases}$$

- $\|w\|_0 = \#$[число ненулевых весов]

Реализация:

- sklearn.linear_model.OrthogonalMatchingPursuit.
- Пример использования.

Orthogonal matching pursuit: метод

1. Инициализировать модель, равную константному нулю.
2. Повторять, пока $\|w\|_0 < K$ (или пока $\|Xw - Y\|_2^2 > \varepsilon$)
   1. добавить признак, максимально коррелирующий с ошибками прогноза последней модели.
   2. переобучить линейную регрессию на данных (отобранные признаки, ошибки прогнозирования)
   3. обновить ошибки прогнозирования

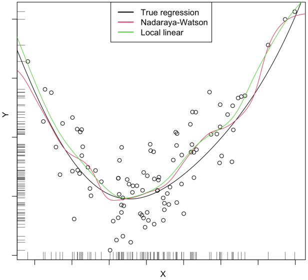
- Метод обобщается
  - на др. меру взаимосвязи признаков и откликов
  - на др. алгоритм прогнозирования (корреляция-только с линейными)

Локальная линейная регрессия

Вместо локальной константы можно оптимизировать локально линейную регрессию:

$$\sum_{i=1}^{N} \alpha_i(x)(x_i^T w - y_i)^2 \to \min_{w \in \mathbb{R}}; \quad \widehat{y}(x) = x^T w$$

Она устойчивее, лучше аппроксимирует области низкой плотности объектов, но вычислительно сложнее.



Заключение

- Лин. регрессия - интерпретируемое аналитическое решение.
- Нелинейные закономерности моделируются:
  - добавлением нелинейных преобразований признаков
  - использованием нелинейной функции $f_w(x)$
- Регуляризация позволяет:
  - считать прогнозы для линейно-зависимых признаков
  - плавно настраивать сложность модели
  - отбирать признаки (лассо регрессия)
- Orthogonal matching pursuit также отбирает признаки.
- Различные функции потерь приводят к разным прогнозам.
- Устойчивость к выбросам достигается:
  - настройкой по модулям (а не квадратам) ошибок
  - взвешенным учётом наблюдений (робастная регрессия)