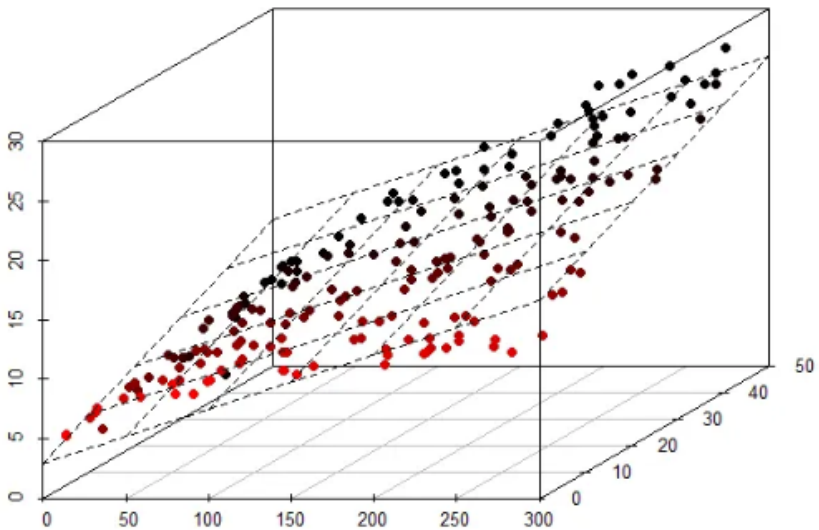
Линейная регрессия и обобщения

Виктор Китов
victorkitov.github.io

Курс поддержан фондом 'Интеллект'



Победитель конкурса VK среди курсов по IT

Содержание

Линейная регрессия

- Линейная регрессия

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

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

- Если смещение $\widehat{\beta}_0$ явно не указано, всегда включают константный признак в x.
- Предположения:
  - каждый x^i линейно влияет y с коэффициентом $\widehat{\beta}_i$
  - 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!)
  - $\widehat{\beta}$ 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

Определим $X \in \mathbb{R}^{N \times D}$, $\{X\}_{ij}$ - значение j-го признака i-го объекта, $Y \in \mathbb{R}^N$, $\{Y\}_i$ - отклик i-го объекта.
Метод наименьших квадратов (МНК, ordinary least squares):

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

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

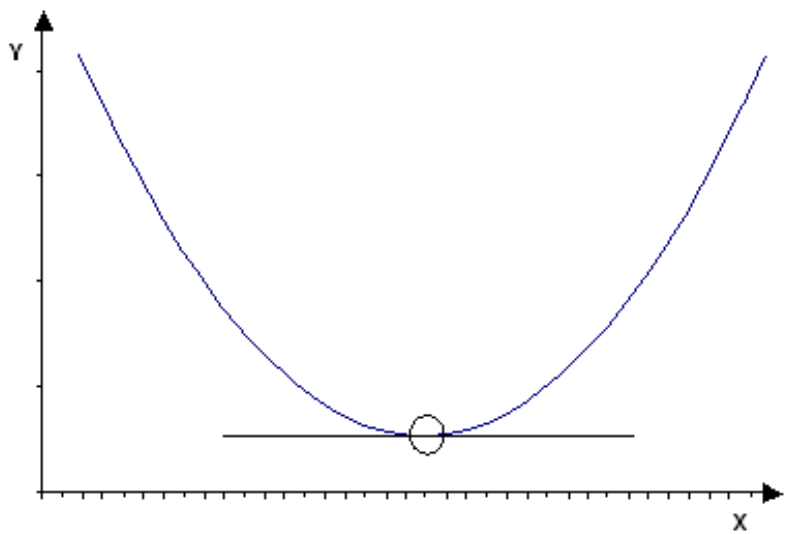
$$\left(\sum_{n=1}^{N} x_n x_n^T \right) \widehat{\beta} = \sum_{n=1}^{N} x_n y_n$$

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

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

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.



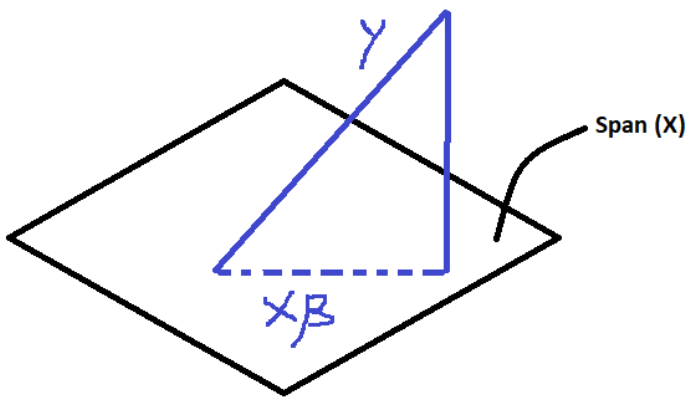
[1] Будет ли суперпозиция произвольных выпуклых ф-ций выпуклой?

Геометрическая интерпретация

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

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

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



Линейно зависимые признаки - проблема

- Решение $\widehat{\beta} = (X^T X)^{-1} X^T Y$ 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 $\widehat{\beta}$ для зависимых признаков:
    - линейная зависимость: $\exists \alpha:\ x^T \alpha = 0\, \forall x$
    - предположим $\widehat{\beta}$ - решение $\sum_{n=1}^{N} \left(x_n^T \beta - y_n\right)^2 \to \min_\beta$
    - тогда $\widehat{\beta} + k\alpha$ - тоже решение $\forall k \in \mathbb{R}: x^T \widehat{\beta} \equiv x^T \widehat{\beta} + k x^T \alpha \equiv x^T (\widehat{\beta} + k\alpha)$.
- При почти зависимых признаках ($X^T X$ плохо обусловлена, т.е. $\lambda_{max}/\lambda_{min}$ велико):
  - $\widehat{\beta}$ неустойчиво и принимает большие по модулю значения.

Линейно зависимые признаки - решение

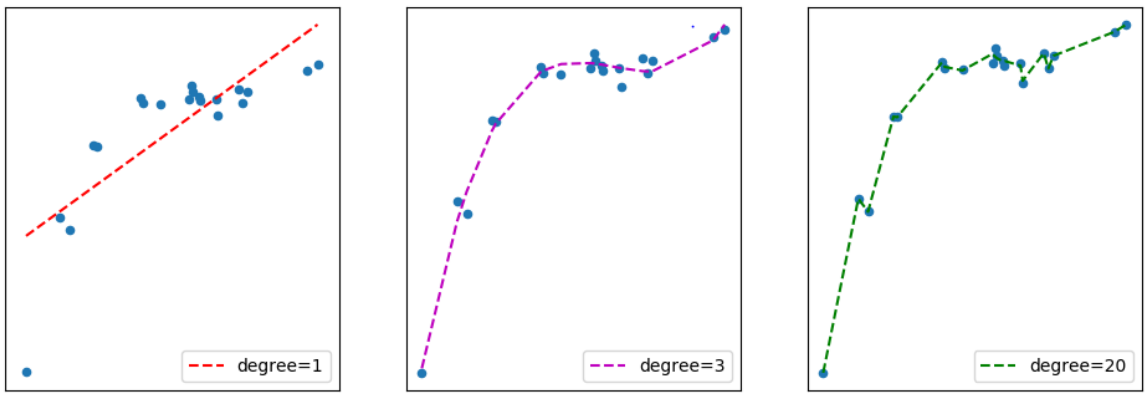
- Проблема может быть решена:
  - отбором признаков (feature selection)
  - снижением размерности (dimensionality reduction)
  - накладыванием доп. условий на решение (регуляризация)
    - $\|\beta\|$ должна быть мала
    - некоторые β_i должны быть неотрицательные
    - ...

Нелинейные зависимости в линейной регрессии

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

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

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



Underfit
High Bias

Correct Fit
Low Bias

Overfit
Low Bias

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 $\widehat{y} = f(x|\beta)$

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

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

- В общем случае не существует аналитического решения $\widehat{\beta}$.
  - 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 $R(\beta)$:

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

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

  $R(\beta) = ||\beta||_1$, Лассо регрессия (Lasso regression)
  $R(\beta) = ||\beta||_2^2$ Гребневая регрессия (Ridge regression)

- 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

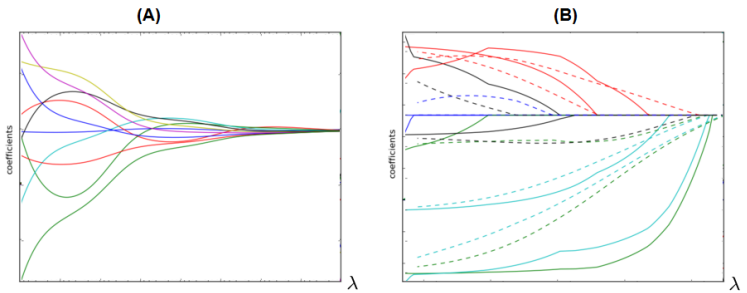
```
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_absolute_error

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 $\widehat{\beta}$ от λ

- Зависимость $\widehat{\beta}$ 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 $\|\beta\|$.

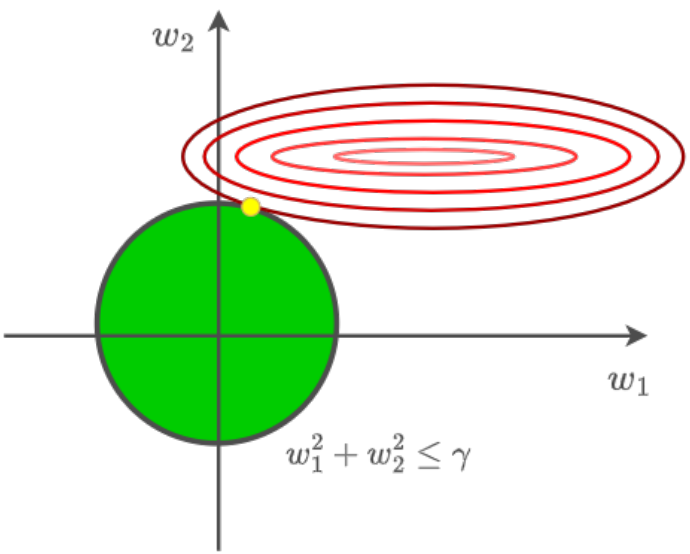
Разное поведение L1 и L2 регуляризации

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

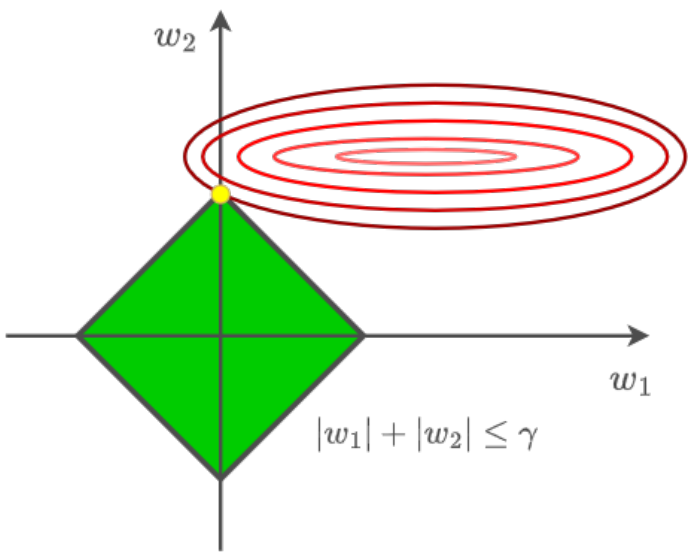
$$L(w)+\lambda R(w)\to \min_w \iff \begin{cases} L(w)\to \min_w \\ R(w)\le \gamma \end{cases}$$

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:

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

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

- 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

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

Условие стационарности (равенство нулю производной):

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

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

$$\left(X^TX + \lambda I\right)\widehat{\beta} = X^TY$$

поэтому

$$\widehat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

$X^TX + \lambda I$ всегда невырождена как сумма $X^TX \succeq 0$ 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:

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

Учет разных признаков с разной силой

- При масштабированию признаков прогнозы лин. регрессии

Учет разных признаков с разной силой

- При масштабированию признаков прогнозы лин. регрессиине изменятся:

$$\widehat{y} = \widehat{\beta}_1 x^1 + \widehat{\beta}_2 x^2 + ... \overset{x^1 \to x^1/\alpha}{\longrightarrow} \left(\alpha\widehat{\beta}_1\right)\left(\frac{x^1}{\alpha}\right) + \widehat{\beta}_2 x^2 + ...$$

- А с регуляризацией изменятся:

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

- 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]

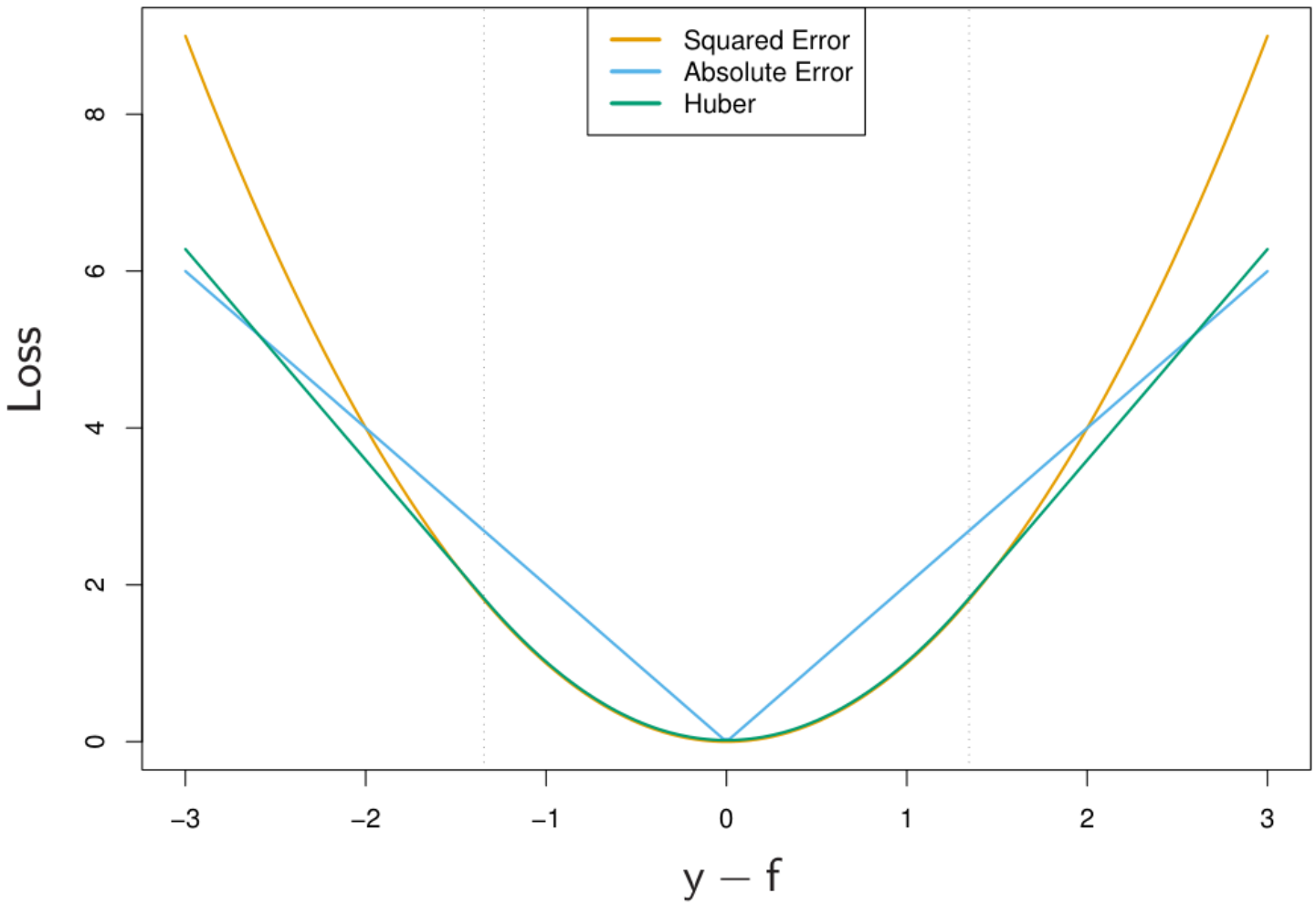
- Обобщим квадратичные потери на произвольные:

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

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:

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

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

Оптимальный прогноз для абсолютной ошибки

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

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

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

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

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

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

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

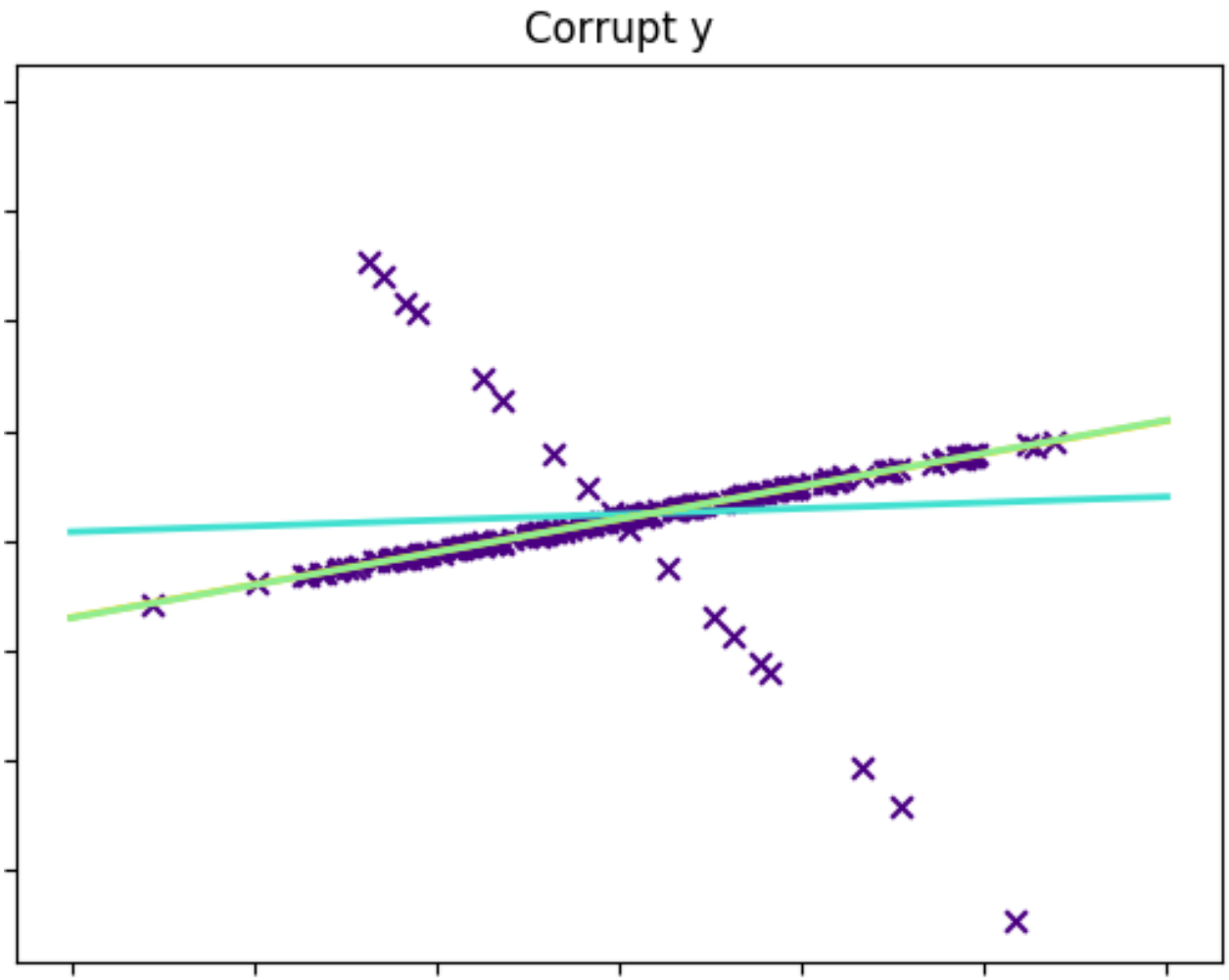
$$\sum_{n=1}^{N} w_n (x_n^T \beta - y_n)^2 \to \min_{\beta \in \mathbb{R}^D}$$

$$w_1 \geq 0, \ldots w_N \geq 0$$

- 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 $w_1 = ... = w_N = 1/N$
- Повторять до сходимости:
  - оценить регрессию $\widehat{y}(x)$ используя (x_i, y_i) с весами w_i.
  - для каждого $i = 1, 2, ...N$:
    - переоценить $\varepsilon_i = \widehat{y}(x_i) - y_i$
    - пересчитать веса $w_i = K\left(|\varepsilon_i|\right)$
  - нормализовать веса $w_i = \frac{w_i}{\sum_{n=1}^{N} w_n}$

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:

$$\begin{cases} \|X\beta - Y\|_2^2 \to \min_\beta \\ \|\beta\|_0 \leq K \end{cases}$$

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

$$\begin{cases} \|\beta\|_0 \to \min_\beta \\ \|X\beta - Y\|_2^2 \leq \varepsilon \end{cases}$$

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

Orthogonal matching pursuit: метод

1. Инициализировать модель, равную константному нулю.
2. Повторять, пока $\|\beta\|_0 < K$ (или пока $\|X\beta - Y\|_2^2 > \varepsilon$)
   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)

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:
  - применением L_1 потерь (лассо регрессия)
  - взвешенным учётом наблюдений (робастная регрессия)