Линейные модели. Регрессия

Формальная постановка

Ищем решающую функцию в виде

$$y = F(w, x) = w^T x,$$

где w — настраиваемые веса, x — признаки.
Решение:

$$w^* = \arg\min L\left(w^T X, y\right),$$

где (X, y) — обучающая выборка.

Можем влиять на:

- x
- L

Метод наименьших квадратов

Решение:

$$w^* = \arg\min \|w^T X - y\|,$$

Если L_2-норма, то

$$w^* = \arg\min \|w^T X - y\|_2^2 = \arg\min (w^T X - y)(w^T X - y),$$

$$\frac{d}{dw}(w^T X - y)(w^T X - y) = 0,$$

откуда

$$w^* = \left(X^T X\right)^{-1} X^T y.$$

Вероятностная интерпретация

Пусть

$$y = w^T x + e,$$

где e — случайный шум с нулевым средним $e \sim N(0, \sigma^2)$. Тогда y также случайная величина и её правдоподобие при параметрах w

$$p(y|x, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y - w^T x)^2}{\sigma^2}}.$$

Можно максимизировать правдоподобие всей выборки

$$\arg\max_w \log p(y|X, w) = \arg\min_w \sum_i \left(y_i - w^T x_i\right)$$

и пользоваться методами теории вероятностей и статистики.

Теорема Гаусса-Маркова

Theorem

Если ошибки модели $y = w^T X + e$

1. *некореллированы* $cov(e_i, e_j) = 0 \ \forall i \neq j$,
2. *имеют одинаковую дисперсию* $var(e_i) = \sigma^2$,
3. *и нулевое среднее* $E(e_i) = 0$,

то модель обладает наименьшим разбросом из всех несмещенных линейных решений

Преобразования над признаками

Какие x бывают:

- непосредственно признаки: $x \in \mathbb{R}^n$,
- мономы: $u \in \mathbb{R}^n$, $x = \prod u_i$,
- произвольные функции: $u \in \mathbb{R}^n$, $x = f(u)$.

Мультиколлинеарность

$$w^* = \left(X^T X\right)^{-1} X^T y.$$

Нужно находить псевдообратную матрицу. Всегда ли это возможно и/или надёжно?

Мультиколлинеарность — наличие линейной зависимости между факторами регрессионной модели. Приводит к плохо обусловленной матрице $\left(X^T X\right)$.

$$\frac{\left\|(A + \Delta A)^{-1} - A^{-1}\right\|}{\|A^{-1}\|} \leq k(A) \frac{\|\Delta A\|}{\|A\|}.$$

Улучшение модели

Для заданных признаков и целевой функции разброс меньше не сделать (теорема Гаусса-Маркова). Значит будем менять целевую функцию (МНМ) или вводить ограничения:

$$\arg\min_w \|w^T X - y\|$$
$$R(w) \le \rho$$

или

$$\arg\min_w R(w)$$
$$\|w^T X - y\| \le \varepsilon$$

или

$$\arg\min_w \|w^T X - y\| + \lambda R(w).$$

Байесовская интерпретация

Ввели априорное распределение на решения:

$$w^* = \arg\max_w P(y|w^T X)P(w)$$

эквивалентно

$$w^* = \arg\max_w \sum_i \log P(y_i|w^T x_i) + \log P(w)$$

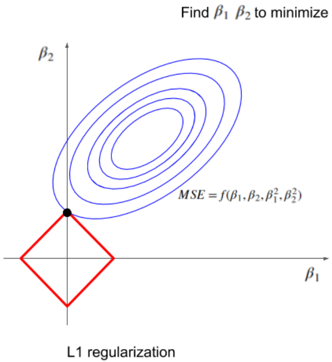
эквивалентно

$$w^* = \arg\min_w \|w^T X - y\| - z \log P(w)$$

Используемые регуляризации

- $\|w\|_0$ — принцип минимальной длины описания (best subset regression). $\|w\|_0$ — число ненулевых элементов w. NP-трудная задача.
- $\|w\|_1$ — LASSO (least absolute shrinkage and selection operator). Априорное распределение Лапласа).
- $\|w\|_2$ — Ridge-регрессия, регуляризация Тихонова. Априорное нормальное распределение. Решение можно получить аналитически.

Геометрия регуляризации



Преимущества и недостатки линейных моделей

Плюсы:

- Простота обучения и использования;
- быстро работает;
- интерпретируемость;
- можно пользоваться статистикой и что-то там доказывать;
- нормально работают, когда мало данных;
- не склонны к переобучению.

Недостатки:

1. Может быть слишком простой для вашей зависимости;
2. может плохо работать, если забыть/не суметь отмасштабировать признаки.