

# 1 A3

## 1.1 $L_2$ -регуляризация

В качестве некоторой эфристики рассмотрим введение дополнительного штрафа к «обучению» в виде

$$R(w) = \tau \|w\|^2 = \tau w^T w, \quad \Rightarrow \quad \mathcal{L} = \|Xw - y\|^2 + \tau \|w\|^2.$$

Всё также есть выпуклая функция, так что её экстремум

$$\partial_w \mathcal{L} = 2X^T(Xw - y) + 2\tau w = 0, \quad \Rightarrow \quad X^T X + \tau \cdot \mathbb{1} = X^T y, \quad \Rightarrow \quad w = (X^T X + \tau \cdot \mathbb{1})^{-1} X^T y.$$

Остается только определиться с выбором параметра  $\tau$ , которая происходит через кроссвалидацию.

## 1.2 $L_1$ -регуляризация

В качестве альтернативной эвристики можно рассмотреть  $L_1$ -регуляризацию:

$$R(w) = \mu \|w\|_1 = \mu \sum_{\alpha=1}^F |w_\alpha|, \quad \Rightarrow \quad \mathcal{L} = \|Xw - y\|^2 + \mu \sum_{\alpha} |w_\alpha|.$$

Теперь функцию потерь не аналитична. Теперь минимум может располагаться на изломах, что будет соответствовать обнулению некоторых признаков. То есть  $L_1$ -регуляризация – отбор признаков.

Можно совместить  $L_2 + L_1$ , тогда получится ElasticNet, что в среднем работает лучше.

## 1.3 Гауссова вероятностная модель

Рассмотрим некоторую модель

$$y = Xw + \varepsilon, \quad \Leftrightarrow \quad y_i = X_{i\alpha} w_\alpha + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}[0, \sigma^2].$$

Для поиска оптимальных весов попробуем применить *принцип максимизации правдоподобия*. Пусть зафиксированы некоторые  $x_i$ , тогда

$$P_w(y) \rightarrow \max_w \quad \Leftrightarrow \quad \log P_w(y) \rightarrow \max_w.$$

Вероятность можно переписать, как

$$P_w(y) = \prod_{i=1}^l \mathcal{N}[0, \sigma^2][\varepsilon_i] = \prod_{i=1}^l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - X_{i\alpha} w_\alpha)}{2\sigma^2}\right),$$

логарифмируя, получаем плюшку:

$$\log P_w(y) = \sum_{i=1}^l \left( -\frac{1}{2} \log 2\pi\sigma^2 - \frac{(y_i - X_{i\alpha} w_\alpha)}{2\sigma^2} \right) \Rightarrow \max_w,$$

где  $(y_i - X_{i\alpha} w_\alpha) \sim \mathcal{L}$ .

То есть можно утверждать, что МНК  $\sim$  максимизация правдоподобия с гауссовой ошибкой:

$$L = -2\sigma^2 \log P_w(y) + \text{const}.$$

Дифференцируя, находим

$$\frac{\partial \log P_w(y)}{\partial \sigma^2} = -\frac{l}{\sigma^2} + \frac{\text{RSS}}{2\sigma^4} = 0, \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{\text{RSS}}{l}.$$

Вообще, можно встретить

$$\hat{\sigma}^2 = \frac{\text{RSS}}{l - F},$$

где  $l$  – объем выборки,  $F$  – количество параметров,  $l - F$  – количество степеней свободы.

## 1.4 Вероятностный смысл регуляризации

Пусть также зафиксированы  $x_i$ , а правдоподобие – вероятность пронаблюдать  $y$  при выбранных  $w$ . Тогда

$$P_w(y) = P(y|w, x), \quad P(A|B) = \frac{P(A, B)}{P(B)}.$$

Можно модифицировать эту схему, введя априорное распределение  $p_\gamma(w)$ , где  $\gamma$  – гиперпараметр, тогда  $w$  – случайная величина. В таких моделях обычно максимизируют правдоподобие данных + модели:

$$\tilde{P}(y, w|x) \rightarrow \max_w, \quad \tilde{P}(y, w|x) = \underbrace{p(y|x, w)}_{P_w(y)} \cdot \underbrace{p_\gamma(w)}_{\text{апр. по } w}.$$

Отсюда найдём, что

$$\log \tilde{P} = \log P + \log p_\gamma(w) = \text{const} - \frac{1}{2\sigma^2} \mathcal{L} + \log p_\gamma(w).$$

Занятно, при  $p_\gamma$  – гауссово, получим  $L_2$ -регуляризацию:

$$p_\gamma(w_\alpha) = \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{w_\alpha^2}{2\gamma}\right), \quad \Rightarrow \quad \log p_\gamma = \text{const} - \frac{w_\alpha^2}{\gamma},$$

и получившийся параметр регуляризации –  $\tau = \frac{\sigma^2}{\gamma}$ .

Альтернативный вариант – распределение  $p_\gamma$  по Лапласу:

$$p_{\tilde{\gamma}} = \frac{1}{2\tilde{\gamma}} \exp\left(-\frac{|w_\alpha|}{\tilde{\gamma}}\right), \quad \Rightarrow \quad \mu = \frac{2\sigma^2}{\tilde{\gamma}},$$

то есть получили  $L_1$ -регуляризацию.

Подчеркнем, что регуляризация эквивалентна некоторому априорному распределению.

## 1.5 Погрешность весов

Вспомним теорему Байеса:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{\sum_A P(B|A) \cdot P(A)}.$$

Действительно,

$$P(A, B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Также можно записать, что

$$P(B) = \sum_A P(A, B) = \sum_A P(B|A) \cdot P(A).$$

Итого: априорное распределение на  $A$  –  $p_\gamma(w)$ ; функция апостериорного распределения на  $A$  –  $p(w|x, y)$ ; и модель  $p(y|w, x) = P(B|A)$ . Тогда

$$p(w|x, y) = \frac{p(y|w, x) \cdot p_\gamma(w)}{\int dw p(y|w, x) p_\gamma(w)}.$$