1 А3

1.1 L_2-регуляризация

В качестве некоторой эфристики рассмотрим введение дополнительного штрафа к «обучению» в виде
$$R(w) = \tau\|w\|^2 = \tau w^{\mathrm{T}} w, \quad \Rightarrow \quad \mathcal{L} = \|Xw - y\|^2 + \tau\|w\|^2.$$
Всё также есть выпуклая функция, так что её экстремум
$$\partial_w \mathcal{L} = 2X^{\mathrm{T}}(Xw - y) + 2\tau w = 0, \quad \Rightarrow \quad X^{\mathrm{T}}X + \tau \cdot \mathbb{1} = X^{\mathrm{T}} y, \quad \Rightarrow \quad w = \left(X^{\mathrm{T}}X + \tau \cdot \mathbb{1}\right)^{-1} X^{\mathrm{T}} y.$$
Остается только определиться с выбором параметра τ, 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 зафиксирован некоторые x_i, тогда
$$P_w(y) \to \max_w \quad \Leftrightarrow \quad \log P_w(y) \to \max_w.$$
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:
$$\log P_w(y) = \sum_{i=1}^{l} \left(-\frac{1}{2} \log 2\pi\sigma^2 - \frac{(y_i - X_{i\alpha} w_\alpha)}{2\sigma^2}\right) \Longrightarrow \max_w,$$
где $(y_i - X_{i\alpha} w_\alpha) \sim \mathcal{L}$.

То есть можно утвеждать, что МНК $\sim$ максимизацияя правдоподобия с гауссовой ошибкой:
$$L = -2\sigma^2 \log P_w(y) + \mathrm{const}.$$
Дифференцируя, находим
$$\frac{\partial \log P_w(y)}{\partial \sigma^2} = -\frac{l}{\sigma^2} + \frac{\mathrm{RSS}}{2\sigma^4} = 0, \quad \Rightarrow \quad \hat{\sigma^2} = \frac{\mathrm{RSS}}{l}.$$
Вообще, можно встретить
$$\hat{\sigma^2} = \frac{\mathrm{RSS}}{l - F},$$
где l – объем выборки, F – количество параметров, $l - F$ – количество степеней свободы.

1.4 Вероятностный смысл регуляризации

Пусть также зафиксированы x_i, а правдоподие – 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 γ – гиперпараметр, тогда w – случайная величина. В таких моделях обычно максимизируют правдоподобие данных + модели:

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

Отсуюда 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_γ – гауссово, получим L_2-регуляризацию:

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

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

Альтернативный вариант – распределение p_γ по Лапласу:

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

то есть получили L_1-регуляризацию.

Подчеркнем, что регуляризация эквивалентна некоторому априорному распредлению.

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

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

2 В1

Рещение СЛАУ

Рассмотрим систему линейный алгебраических уравнений:

$$A\boldsymbol{x} = \bar{b}, \qquad A \equiv \alpha_{ij}.$$

I. Как можно решать? через правило Крамера:

$$x_i = \frac{\Delta_i}{\Delta}, \qquad i = 1,\, 2,\, \ldots,\, n, \qquad \Delta = \det A \neq 0, \qquad \Delta_1 = \det A^i,$$

где A^i – матрица A, в которой заменили i-й столбец на $\boldsymbol{b}$.

Сложность вычисления $O_{\text{det}} = O(n!)$, а ещё мы должны посчитать $n+1$ определитель. Зато можно находить конкретные x_i.

II. Альтернатива: поиск обратной матрицы:

$$A\boldsymbol{x} = \boldsymbol{b}, \qquad \Rightarrow \qquad \boldsymbol{x} = A^{-1}\boldsymbol{b}.$$

Сложность вычислений: $O(n^3)$ или $O(n^2) \cdot O_{\text{det}}$.

III. Метод Гаусса:

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} & b_1 \\ \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} & b_1 \end{pmatrix} \to \begin{pmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1n} & b_1 \\ \\ 0 & 0 & \ldots & \beta_{nn} & b_1 \end{pmatrix},$$

который реализуется элементарными преобразованиями.

Берем и считаем $k_i = \frac{\alpha_{i\rho}}{\alpha_{11}}$, и меняем $a_i = a_i - k_i \cdot a_i$. После обнуления первого столца образуется новая матрица A' размера $n-1$, где мы игнорируем первую строку и первый столбец.

Решение ищем уже через обратный метод Гаусса:

$$x_n = \frac{b_n^*}{\beta_{nn}}, \quad \Rightarrow \quad x_{n-1} = \frac{b_{n-1}^* - b_n^* \frac{\beta_{n-1\ n}}{\beta_{nn}}}{\beta_{n-1\ n-1}}.$$