Задача обучения линейного классификатора

Обучающая выборка: $X^\ell = (x_i, y_i)_{i=1}^\ell, \;\; x_i \in \mathbb{R}^n, \;\; y_i \in \{-1, +1\}$

- Линейная модель классификации:

$$a(x, w) = \operatorname{sign}\langle x, w\rangle$$

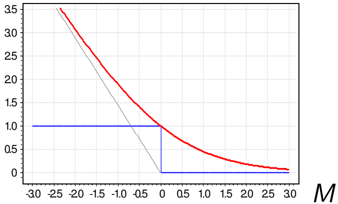
- Непрерывная аппроксимация бинарной функции потерь:

$$Q(w) = \sum_{i=1}^{\ell} \big[a(x_i, w)y_i < 0\big] \leqslant \sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) \to \min_w$$

Отступ (margin) объекта x_i: $\;M_i(w) = \langle x_i, w\rangle y_i$

- Логарифмическая функция потерь, как функция отступа M:

$$\mathscr{L}(M) = \log\big(1 + e^{-M}\big)$$



$(x_i, y_i)_{i=1}^{\ell} \sim p(x, y; w)$ — выборка независимых наблюдений.

Принцип максимума правдоподобия:

$$L(w) = \log \prod_{i=1}^{\ell} p(x_i, y_i; w) = \sum_{i=1}^{\ell} \log P(y_i | x_i; w) p(x_i) \to \max_{w}.$$

Вероятностная модель порождения данных с параметром w:

- $p(x)$ не зависит от параметра модели w,
- $P(y|x; w)$ описывается линейной моделью классификации:

$$P(y_i | x_i; w) = \frac{1}{1 + \exp\bigl(-\langle x_i, w \rangle y_i\bigr)} = \sigma\bigl(\langle x_i, w \rangle y_i\bigr),$$

где $\sigma(M) = \frac{1}{1+e^{-M}}$ — сигмоидная функция.

Тогда задачи $Q(w) \to \min$ и $L(w) \to \max$ эквивалентны:

$$Q(w) = \sum_{i=1}^{\ell} \log\bigl(1 + \exp(-\langle w, x_i \rangle y_i)\bigr) \to \min_{w}.$$

- Метод первого порядка — стохастический градиент:

$$w^{(t+1)} := w^{(t)} + \eta_t \, y_i x_i \, (1 - \sigma_i),$$

η_t — градиентный шаг,
$\sigma_i = \sigma\bigl(\langle x_i, w\rangle y_i\bigr) = \mathsf{P}(y_i|x_i)$ — вероятность правильной классификации x_i.

- Метод второго порядка (Ньютона-Рафсона) приводит к IRLS, Iteratively Reweighted Least Squares:

$$w^{(t+1)} := w^{(t)} + \eta_t (F^\intercal \Lambda F)^{-1} F^\intercal \tilde{y},$$

F — матрица объекты–признаки $\ell \times n$,
$\tilde{y} = \bigl(y_i(1 - \sigma_i)\bigr)$ — модифицированный вектор ответов,
$\Lambda = \mathrm{diag}\bigl((1 - \sigma_i)/\sigma_i\bigr)$ — диагональная матрица.