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:

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

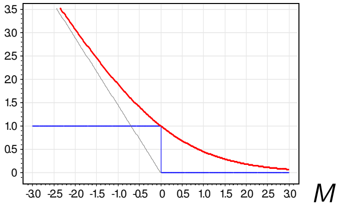
- Непрерывная аппроксимация бинарной функции потерь:

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

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

- Логарифмическая функция потерь, как функция отступа M:

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



Обоснование логарифмической функции потерь

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

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

Оптимизация параметров логистической регрессии

- Метод первого порядка — стохастический градиент:

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

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

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

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

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

Пример. Бинаризация признаков и скоринговая карта

Задача кредитного скоринга:

- x_i — заёмщики
- $y_i \in \{-1(\text{bad}), +1(\text{good})\}$

Бинаризация признаков $f_j(x)$:

$$b_{jk}(x) = [f_j(x) \in D_{jk}]$$

Возраст	до 25	**5**
	25 - 40	**10**
	40 - 50	**15**
	50 и больше	**10**
Собственность	владелец	**20**
	совладелец	**15**
	съемщик	**10**
	другое	**5**
Работа	руководитель	**15**
	менеджер среднего звена	**10**
	служащий	**5**
	другое	**0**
Стаж	1/безработный	**0**
	1..3	**5**
	3..10	**10**
	10 и больше	**15**
Работа_мужа /жены	нет/домохозяйка	**0**
	руководитель	**10**
	менеджер среднего звена	**5**
	служащий	**1**

Пример. Оценивание рисков в кредитном скоринге

Оценка *риска* (математического ожидания) потерь объекта x:

$$R(x) = \sum_{y \in Y} D_{xy}\, \mathsf{P}(y|x) = \sum_{y \in Y} D_{xy} \sigma\bigl(\langle w, x \rangle y\bigr),$$

где D_{xy} — величина потери для (x, y).

Методика VaR (Value at Risk)

Оценка функции распределения потерь:

- для каждого x_i разыгрывается N раз исход $y_i \sim P(y|x_i)$;
- строится эмпирическое распределение потерь $V = \sum_{i=1}^{\ell} D_{x_i y_i}$;
- 99%-квантиль эмпирического распределения определяет величину резервируемого капитала

Регуляризованная логистическая регрессия

- L_2-регуляризация решает проблему мультиколлинеарности (сокращает веса линейно зависимых признаков):

$$Q(w) = \sum_{i=1}^{\ell} \log\bigl(1 + \exp(-\langle w, x_i\rangle y_i)\bigr) + \tau \sum_{j=1}^{n} w_j^2 \to \min_w .$$

- L_1-регуляризация имеет эффект отбора признаков (обнуляет веса w_j неинформативных признаков):

$$Q(w) = \sum_{i=1}^{\ell} \log\bigl(1 + \exp(-\langle w, x_i\rangle y_i)\bigr) + \tau \sum_{j=1}^{n} |w_j| \to \min_w .$$

- Используется также их комбинация — ElasticNet.

Коэффициент регуляризации τ подбирается по скользящему контролю.

Резюме

- *Логистическая регрессия* — это линейный классификатор,
- оценивающий апостериорные вероятности классов $P(y|x)$, необходимые в прикладных задачах оценивания рисков.
- Регуляризация улучшает обобщающую способность логистической регрессии:
  - L_2-регуляризация — при мультиколлинеарности признаков;
  - L_1-регуляризация — для отбора признаков;
  - ElasticNet — для менее агрессивного отбора признаков.