Линейные методы классификации

Виктор Китов
v.v.kitov@yandex.ru

Содержание

Линейная дискриминантная функция

- Классификация среди двух классов ω_1 и ω_2.
- Линейная дискриминантная функция:

$$y(x) = w^T x + w_0$$

- Решающее правило:

$$x \to \begin{cases} \omega_1, & y(x) \geq 0 \\ \omega_2, & y(x) < 0 \end{cases}$$

- Граница классов $B = \{x : y(x) = 0\}$

Свойства

- $x_A, x_B \in B \Rightarrow \begin{cases} y(x_A) = w^T x_A + w_0 = 0 \\ y(x_B) = w^T x_B + w_0 = 0 \end{cases} \Rightarrow$
  $w^T(x_A - x_B) = 0$, поэтому $w \perp B$.
- Расстояние от начала координат до B равно абсолютной величине проекции $x \in B$ на $\frac{w}{\|w\|}$:

$$\langle x, \frac{w}{\|w\|} \rangle = \frac{\langle x, w \rangle}{\|w\|} = \{w^T x + w_0 = 0\} = -\frac{w_0}{\|w\|}$$

- Поэтому $\rho(0, B) = \frac{|w_0|}{\|w\|}$, и w_0 определяет смещение.

Расстояние от x до B

Обозначим через $x_\perp$ вектор проекции x на B, а $r = \langle \frac{w}{\|w\|}, x - x_\perp \rangle$ - проекцию x на B:

$$x = x_\perp + r\frac{w}{\|w\|}$$

Умножим на w и прибавим w_0:

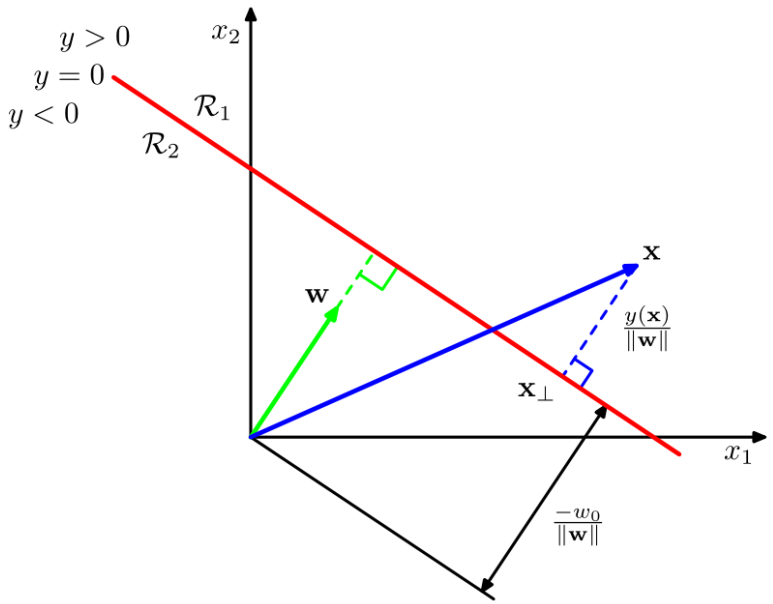
$$w^T x + w_0 = w^T x_\perp + w_0 + r\frac{\langle w, w\rangle}{\|w\|}$$

Используя $w^T x + w_0 = y(x)$ и $w^T x_\perp + w_0 = 0$, получим:

$$r = \frac{y(x)}{\|w\|}$$

Следовательно, с одной стороны гиперплоскости $r > 0 \Leftrightarrow y(x) > 0$, а с другой $r < 0 \Leftrightarrow y(x) < 0$.

Демонстрация



Линейное решающее правило:

$$\widehat{c}(x) = \begin{cases} \omega_1, & y(x) > 0 \\ \omega_2, & y(x) < 0 \end{cases}$$

Граница классов: $y(x) = 0$,
степень уверенности в классификации: $|y(x)| / \|w\|$.

Содержание

Линейные дискриминантные функции

- Линейная дискриминантная функция: $g(x) = w^T x + w_0$,

$$\widehat{\omega} = \begin{cases} \omega_1, & g(x) \geq 0 \\ \omega_2, & g(x) < 0 \end{cases}$$

- Обозначим классы ω_1 и ω_2 через $y = +1$ и $y = -1$. Решающее правило: $y = \text{sign}\, g(x)$.
- Определим дополнительный признак $x_0 \equiv 1$, тогда $g(x) = w^T x = \langle w, x \rangle$ для $w = [w_0, w_1, ..w_D]^T$.
- Определим отступ $M(x) = g(x)y$
  - $M(x) \geq 0$ <=> объект x правильно классифицирован
  - $|M(x)|$ - уверенность классификатора в прогнозе

Выбор весов

- Цель - оптимизация функции потерь:

$$Q_{accurate}(w|X) = \sum_i \mathbb{I}[M(x_i|w) < 0] \to \min_w$$

- Проблема: стандартные методы оптимизации неприменимы, т.к $Q(w, X)$ разрывна.
- Идея решения: аппроксимировать функцию цены сверху гладкой функцией $\mathcal{L}$:

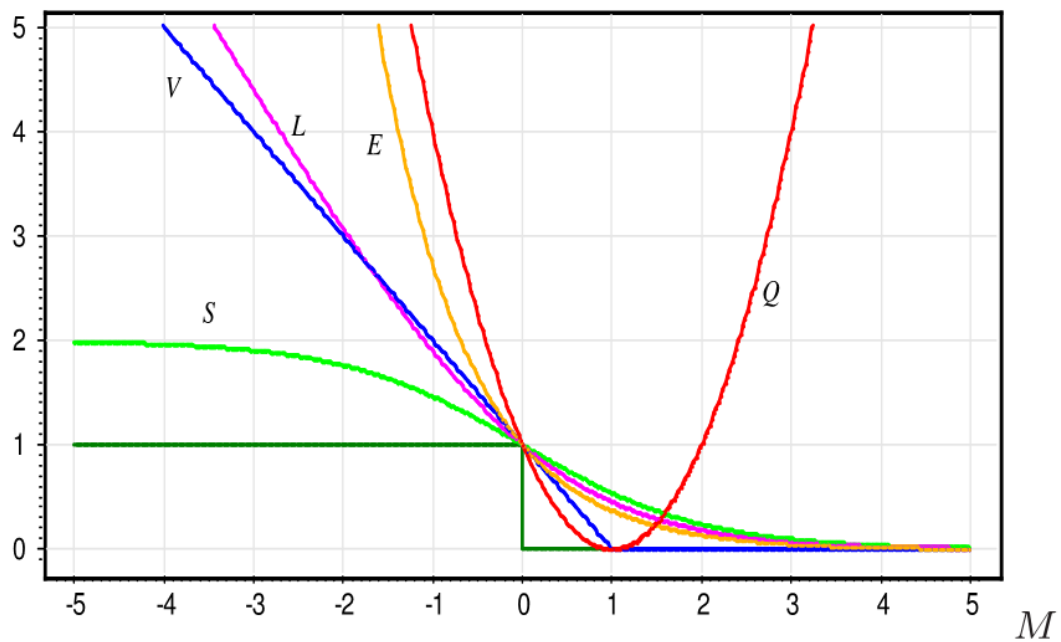
$$\mathbb{I}[M(x_i|w) < 0] \leq \mathcal{L}(M(x_i|w))$$

Аппроксимация целевого критерия

Получаем аппроксимацию эмпирического риска сверху:

$$Q_{accurate}(w|X) = \sum_i \mathbb{I}[M(x_i|w) < 0]$$

$$\leq \sum_i \mathcal{L}(M(x_i|w)) = Q_{approx}(w|X)$$



$$Q(M) = (1-M)^2$$
$$V(M) = (1-M)_+$$
$$S(M) = 2(1+e^M)^{-1}$$
$$L(M) = \log_2(1+e^{-M})$$
$$E(M) = e^{-M}$$

Содержание

Оптимизация

- Оптимизационная задача для получения весов:

$$\begin{aligned} F(w) &= Q_{approx}(w|X,Y) = \sum_{i=1}^{n} \mathcal{L}(M(x_i, y_i|w)) \\ &= \sum_{i=1}^{n} \mathcal{L}(\langle w, x_i \rangle y_i) \to \min_w \end{aligned}$$

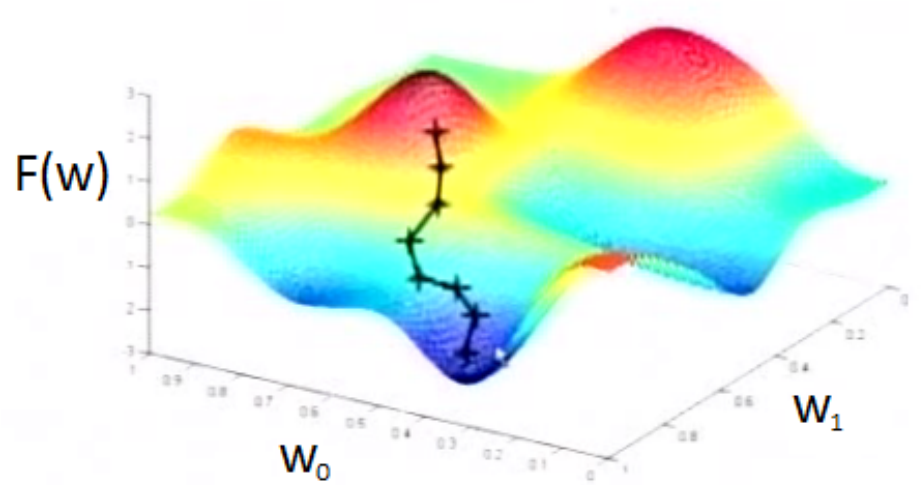
Алгоритм градиентного спуска

ВХОД:
η – параметр, контролирующий скорость сходимости
критерий остановки

АЛГОРИТМ:
инициализировать w_0 случайным образом
пока не выполнен критерий остановки:
$w_{n+1} \leftarrow w_n - \eta \frac{\partial F(w_n)}{\partial w}$
$n \leftarrow n+1$

Алгоритм градиетного спуска

- Критерии остановки:
  - $|w_{n+1} - w_n| < \varepsilon$
  - $|F(w_{n+1}) - F(w_n)| < \varepsilon$
  - $n > n_{max}$
- Субоптимальный метод минимизации в направлении наибольшего уменьшения $F(w)$:



Ускорение сходимости

Метод стохастического градиента

задать начальное приближение w_0
рассчитать $\widehat{Q}_{approx} = \sum_{i=1}^{n} \mathcal{L}(M(x_i|w_0))$
итеративно, до сходимости $\widehat{Q}_{approx}$:

1. выбрать случайное наблюдение (x_i, y_i)
2. пересчитать веса: $w_{n+1} \leftarrow w_n - \eta_n \mathcal{L}'(\langle w_n, x_i \rangle y_i) x_i y_i$
3. оценить ошибку: $\varepsilon_i = \mathcal{L}(\langle w_{n+1}, x_i \rangle y_i)$
4. пересчитать оценку $\widehat{Q}_{approx} = (1-\alpha)\widehat{Q}_{approx} + \alpha\varepsilon_i$
5. $n \leftarrow n+1$

Выбор начальных весов

- $w_0 = w_1 = ... = w_D = 0$
- Для логистической ф-ции $\mathcal{L}$ (из-за асимптоты слева):
  - случайно на интервале $[-\frac{1}{2D}, \frac{1}{2D}]$
- Для др. ф-ции:
  - случайно на произвольном интервале
- $w_i = \frac{\langle x^i, y \rangle}{\langle x^i, x^i \rangle}$

Обсуждение метода

Преимущества

- Легко реализовать
- Работает в online-режиме
- Небольшого подмножества обучающих объектов может быть достаточно для точной оценки

Обсуждение метода

Преимущества

- Легко реализовать
- Работает в online-режиме
- Небольшого подмножества обучающих объектов может быть достаточно для точной оценки

Недостатки

- Субоптимальность - сходимость к локальному оптимуму
- Необходимость выбора η_n:
  - при слишком больших-расходимость
  - при слишком маленьких-медленная сходимость
- Возможно переобучение для больших D и малых N
- для логистической аппроксимации (и всегда, когда $\mathcal{L}(u)$ имеет горизонтальные асимптоты), алгоритм может «застрять» для больших значений $\langle w, x_i \rangle$.

Примеры

Дельта-правило $\mathcal{L}(M) = (M-1)^2$

$$w \leftarrow w - \eta(\langle w, x_i\rangle - y_i)x_i$$

Это также подходит для регрессии и $f(x) = \langle w, x\rangle$ для ф-ции цены $(\langle w, x\rangle - y)^2$, $y \in \mathbb{R}$

$\mathcal{L}(M) = [-M]_+$

Персептрон Розенблатта

$$w \leftarrow w + \begin{cases} 0, & \langle w, x_i\rangle y_i \geq 0 \\ \eta x_i y_i & \langle w, x_i\rangle y_i < 0 \end{cases}$$

Логичное правило, но не пытается расширить полосу разделения между классами.

Содержание

Переобучение

- Ранняя остановка
  - остановка, когда качество перестает улучшаться
- Регуляризация
  - Штраф за большие веса:

$$Q_{approx}^{regularized}(w) = Q_{approx}(w) + \frac{\tau}{2}|w|^2$$

- Шаг градиентного спуска: $w \leftarrow w(1 - \eta\tau) - \eta Q'_{approx}(w)$

Регуляризация

- Удобный прием для контроля сложности модели:

$$Q^{regularized}(w) = Q(w) + \tau||w||_2^2$$

$$Q^{regularized}(w) = Q(w) + \tau||w||_1$$

$$||w||_1 = \sum_{d=1}^{D} |w^d|, \quad ||w||_2 = \sqrt{\sum_{d=1}^{D} (w^d)^2}$$

- Регрессия, оцениваемая методом наименьших квадратов с регуляризацией:
  - $\tau||w||_1$ - LASSO
  - $\tau||w||_2^2$ - Ridge
  - $\alpha||w||_1+\beta||w||_2$ - elastic net:

L_1 норма

- $||w||_1$ регуляризация производит отбор признаков.
- Рассмотрим

$$Q(w) = \sum_{i=1}^{N} \mathcal{L}_i(w) + \lambda \sum_{d=1}^{D} |w_d|$$

- При $\lambda > \sup_w \left| \frac{\partial \mathcal{L}(w)}{\partial w_i} \right|$ становится заведомо лучше положить $w_i = 0$
- Для более высоких λ больше коэффициентов становятся равными нулю.

Содержание

Логистическая регрессия

- Добавим в x константный признак и w_0 к w.
- Сигмоидная функция активации $\sigma(z) = \frac{1}{1+e^{-z}}$.
- Двухклассовая классификация:

$$score(\omega_1|x) = w^T x$$

$$p(\omega_1|x) = \sigma(w^T x)$$

- Многоклассовая классификация:

$$\begin{cases} score(\omega_1|x) = w_1^T x \\ score(\omega_2|x) = w_2^T x \\ \dots \\ score(\omega_C|x) = w_C^T x \end{cases}$$

Логистическая регрессия

Вероятности классов аппроксимируются через soft-max функцию:

$$p(\omega_c|x) = \frac{exp(w_c^T x)}{\sum_i exp(w_i^T x)}$$

w_c, $c = 1, 2, ...C$ определены с точностью до сдвига на произвольный вектор v:

$$\frac{exp((w_c - v)^T x)}{\sum_i exp((w_i - v)^T x)} = \frac{exp(-v^T x)exp(w_c^T x)}{\sum_i exp(-v^T x)exp(w_i^T x)} = \frac{exp(w_c^T x)}{\sum_i exp(w_i^T x)}$$

Обычно сдвигают все w_c на $v = w_C$.

Замечание: нелинейное преобразование score в вероятность могло быть определено и по-другому - получили бы другой метод.

Логистическая регрессия

- Пусть γ_1, γ_2 - цены неправильной классификации классов ω_1 и ω_2.
- Предположим

$$\ln\left(\frac{\gamma_1 p(\omega_1|x)}{\gamma_2 p(\omega_2|x)}\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$$

- это эквивалентно

$$\begin{aligned} p(\omega_2|x) &= \frac{1}{1+exp(\beta_0' + \boldsymbol{\beta}^T \boldsymbol{x})} \\ p(\omega_1|x) &= \frac{exp(\beta_0' + \boldsymbol{\beta}^T \boldsymbol{x})}{1+exp(\beta_0' + \boldsymbol{\beta}^T \boldsymbol{x})} \end{aligned}$$

- где $\beta_0' = \beta_0 - \ln(\gamma_1/\gamma_2)$

Логистическая регрессия

Решающее правило (следуя Байесовскому правилу минимальной цены):

$$x = \begin{cases} \omega_1, & \beta_0' + \boldsymbol{\beta}^T \boldsymbol{x} > 0 \\ \omega_2, & \beta_0' + \boldsymbol{\beta}^T \boldsymbol{x} < 0 \end{cases}$$

Оценка $\beta_0', \boldsymbol{\beta}$ методом максимального правдоподобия:

$$\prod_{i=1}^{N} p(c_i|x_i) \to \max_{\beta_0', \boldsymbol{\beta}}$$

где c_i - класс объекта x_i.

Многоклассовая логистическая регрессия

- Предположение:

$$\ln\left(\frac{\gamma_s p(\omega_s|x)}{\gamma_C p(\omega_C|x)}\right) = \beta_{s0} + \beta_s^T \boldsymbol{x}, \quad s = 1, 2, ...C-1$$

- Вероятности классов (дающие эквивалентное определение):

$$p(\omega_s|x) = \frac{exp(\beta'_{s0} + \beta_s^T x)}{1 + \sum_{s=1}^{C-1} exp(\beta'_{s0} + \beta_s^T x)}, \quad s = 1, 2, ...C-1$$

$$p(\omega_C|x) = \frac{1}{1 + \sum_{s=1}^{C-1} exp(\beta'_{s0} + \beta_s^T x)}$$

$$\beta'_{s0} = \beta_{s0} - \ln(\gamma_s/\gamma_C)$$

- Интерпретация: soft-max от дискриминатных функций (для классов $\omega_1, \omega_2, ...\omega_{C-1}$) и константы (для класса ω_C).

Многоклассовая логистическая регрессия

- Решающее правило (Байесовское правило минимальной ожидаемой цены):
- $c = \arg\max_c \beta_{c0} + \beta_c^T x$, если $\beta_{c0} + \beta_c^T x > 0$ иначе сопоставить x классу C.
- Оценивание методом максимального правдоподобия:

$$\prod_{i=1}^{N} p(c_i|x_i) \to \max_{\beta_0', \beta}$$

Функция цены

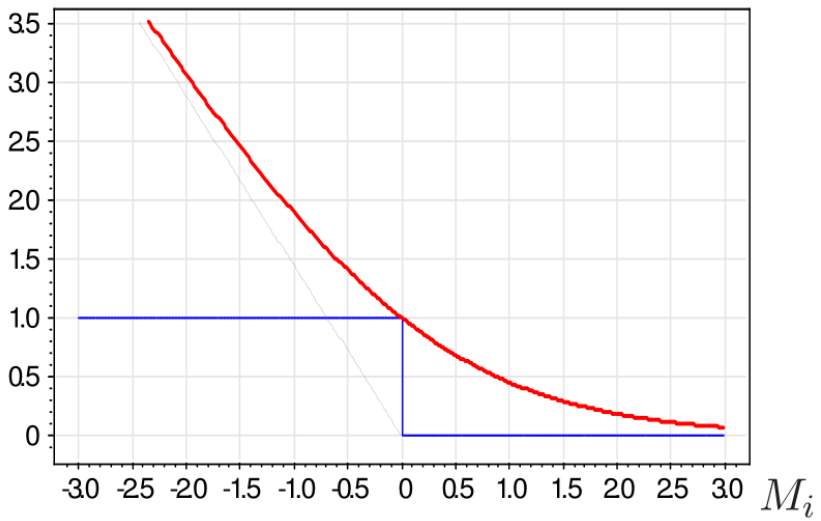
Для 2-х классов $p(y|x) = \sigma(\langle w, x\rangle y)$, где $\sigma = \frac{1}{1+e^{-z}}$, $w = [\beta_0', \beta]$, $x = [1, x_1, x_2, ...x_D]$.

Оценка методом максимального правдоподобия:

$$\prod_{i=1}^{N} \sigma(\langle w, x_i\rangle y_i) \to \max_w$$

эквивалентна

$$\sum_{i=1}^{N} \ln(1 + e^{-\langle w, x_i\rangle y_i}) \to \min_w$$



Следовательно, мажорирующая ф-ция для логистической регрессии $\mathcal{L}(M) = \ln(1 + e^{-M})$.