Линейные методы классификации I

Виктор Владимирович Китов
v.v.kitov@yandex.ru

МГУ им.Ломоносова, ф-т ВМиК, кафедра ММП.

I семестр 2015 г.

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:

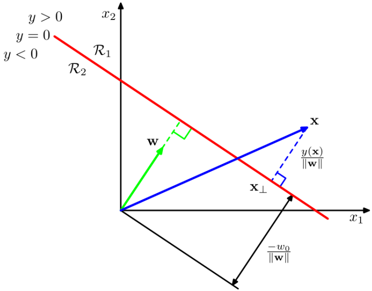
$$w^T x + w_0 = w^T x_\perp + w_0 + r\frac{\langle w, w\rangle}{\|w\|}$$

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: $|y(x)| / \|w\|$.

Неоднозначность многоклассовой классификации

Предположим, нужно классифицировать среди 3х классов C_1, C_2, C_3.

Схема «один против всех»:

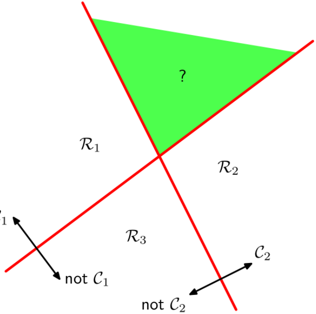
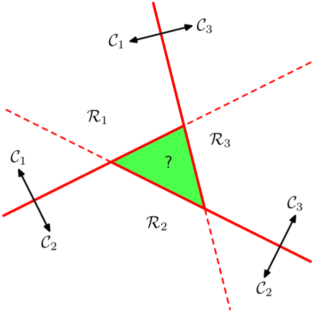


Схема «один против одного»:



Подход «один против всех с весами»

- Классификация среди $\omega_1, \omega_2, \ldots \omega_C$.
- Используем C дискриминантных функций:

$$g_c(x) = w_c^T x + w_{c0}, \quad c = 1, 2, \ldots C.$$

- Решающее правило:

$$\widehat{c}(x) = \arg\max_c g_c(x)$$

- Граница между классами ω_i и ω_j линейна:

$$\left(w_i - w_j\right)^T x + \left(w_{i0} - w_{j0}\right) = 0$$

Выпуклость областей прогнозируемых классов

Области прогнозируемых классов выпуклы.

Доказательство: Допустим $\widehat{c}(x_A) = \widehat{c}(x_B) = c, \quad c = 1, 2, ...C,$ что по определению означает

$$w_c^T x_A + w_{c0} \geq w_k^T x + w_{k0} \quad \forall k \neq c \qquad (1)$$
$$w_c^T x_B + w_{c0} \geq w_k^T x + w_{k0} \quad \forall k \neq c \qquad (2)$$

Для $\lambda x_A + (1-\lambda)x_B$, $\lambda \in (0,1)$ суммируя **(??)** и **(??)** с весами λ и $(1-\lambda)$, получим:

$$w_c^T(\lambda x_A + (1-\lambda)x_B) + w_{c0} \geq w_k^T(\lambda x_A + (1-\lambda)x_B) + w_{k0} \quad \forall k \neq c$$

что означает, что прогноз $\widehat{c}(\lambda x_A + (1-\lambda)x_B) = c$, и область каждого класса $c = 1, 2, ...C$ выпукла.

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: $y = \operatorname{sign} g(x)$.
- 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:

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

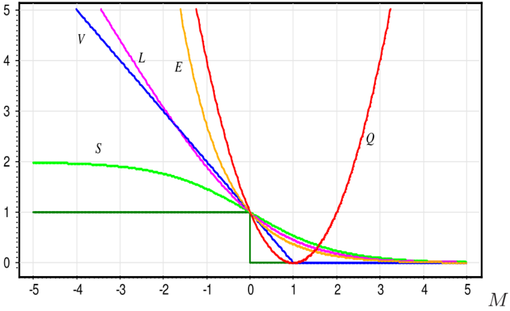
- Проблема: стандартные методы оптимизации неприменимы, т.к $Q(w, X)$ разрывна.
- Идея решения: аппроксимировать функцию цены сверху гладкой функцией $\mathcal{L}$:

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

Аппроксимация целевого критерия

Получаем аппроксимацию эмпирического риска сверху:

$$
\begin{aligned}
Q_{accurate}(w|X) &= \sum_i \mathbb{I}[M(x_i|w) < 0] \\
&\leq \sum_i \mathcal{L}(M(x_i|w)) = Q_{approx}(w|X)
\end{aligned}
$$



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

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:

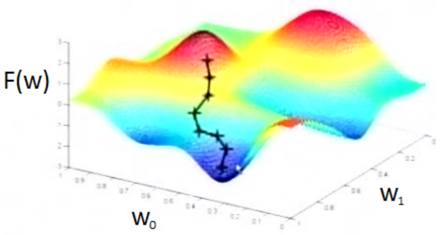
η — параметр, контролирующий скорость сходимости
критерий остановки

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

- $w_0 = w_1 = \dots = w_D = 0$
- 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 $(\langle w, x \rangle - y)^2$, $y \in \mathbb{R}$

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

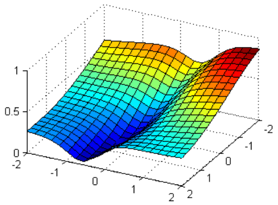
Персептрон Розенблатта

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

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

Рекомендации к использованию

- Сходимость быстрее для нормализованных признаков
  - нормализация частично решает проблему «вытянутых долин»



  - для $\mathcal{L}$ с левыми горизонтальными асимптотами: $\langle w, x_i \rangle y_i$ ограничено на начальных итерациях, и метод не «застревает»

Рекомендации к использованию

- Быстрее сходимость, когда больше совершается ошибок:
  - сэмплировать наблюдения с вероятностями, пропорциональными $\varepsilon_i = \mathcal{L}(\langle w, x_i \rangle y_i)$
    - ускорение вычислений: пересчитывать w (и связанные $\mathcal{L}(\langle w, x_i \rangle y_i)$) только когда $\varepsilon_i \geq \delta$ для некоторого порога $\delta > 0$.
  - повышать разнообразие сэмплируемых объектов
    - например, за счет целенаправленного сэмплирования из разных классов
- Локальный оптимум совпадает с глобальным для выпуклой $\mathcal{L}(w|x, y)$
  - для невыпуклого случая нужно запускать процедуру для различных начальных приближений и выбрать решение, дающее минимальное значение $Q_{approx}(w|X, Y)$

Выбор η

- Больше η => больше риск, что алгоритм будет расходиться.
- Тест для контроля сходимости: смотреть динамику $Q_{approx}(w)$ (или $\widehat{Q}_{approx}(w)$) от номера итерации t.
- Тип η:
  - $\eta_t = \eta = const$
  - веса, изменяющиеся по фиксированному правилу:
    - условия сходимости метода:

    1. $\eta_t \to 0$
    2. $\sum_{t=1}^{\infty} \eta_t = \infty$
    3. $\sum_{t=1}^{\infty} \eta_t^2 < \infty$
  - Пример: $\eta_t = \frac{1}{t}$

Выбор

- Шаг, зависимый от данных:
  - На каждом шаге $\eta_t = \arg\min_{\eta} Q_{approx}(w - \eta \frac{\partial Q_{approx}}{\partial w})$
  - Часто существует аналитическое решение для η

Содержание

Оценка методом максимального правдоподобия

- $X = \{x_1, x_2, ...x_n\}$, $Y = \{y_1, y_2, ...y_n\}$ - обучающая выборка, $(x_i, y_i) \sim p(y|x, w)$
- Оценка максимального правдоподобия $\widehat{w} = \arg\max_w p(Y|X, w)$
- Используя предположение о независимости $y_i|x_i$ $i = 1, 2, ...N$:

$$\prod_{i=1}^{N} p(y_i|x_i, w) = \sum_{i=1}^{N} \ln p(y_i|x_i, w) \to \max_w$$

- Оптимизация цены:

$$\sum_{i=1}^{n} \mathcal{L}(g(x_i)y_i|w) \to \min_w$$

- Взаимосвязь двух методов:

$$\mathcal{L}(g(x_i)y_i|w) = -\ln p(y_i|x_i, w)$$

Максимальная апострериорная вероятность

- $X = \{x_1, x_2, ...x_N\}$, $Y = \{y_1, y_2, ...y_N\}$ - обучающая выборка из н.о.р. $(x_i, y_i) \sim p(x, y|w)$
- w - не константа, а с.в. $w \sim p(w)$
- Поскольку

$$p(w|X, Y) = \frac{p(X, Y, w)}{p(X, Y)} = \frac{p(X, Y|w)p(w)}{p(X, Y)} \propto p(X, Y|w)p(w)$$

то справедлива оценка *максимизации апостериорной вероятности*:

$$w = \arg\max_w p(w|X, Y) = \arg\max_w p(X, Y|w)p(w)$$

$$\sum_{i=1}^{n} \ln p(x_i, y_i|\theta) + \ln p(w) \to \max_w$$

$\ln p(w)$ соответствует регуляризации.

Варианты априорных вероятностей

- Нормальное распределение

$$\ln p(w, \sigma^2) = \ln\left(\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{||w||_2^2}{2\sigma^2}}\right) = -\frac{1}{2\sigma^2}||w||_2^2 + \mathrm{const}$$

- Распределение Лапласа

$$\ln p(w, C) = \ln\left(\frac{1}{(2C)^n} e^{-\frac{||w||_1}{C}}\right) = -\frac{1}{C}||w||_1 + \mathrm{const}$$

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:

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

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

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

- Регрессия, оцениваемая методом наименьших квадратов с регуляризацией:
  - $\tau||w||_1$ - LASSO
  - $\tau||w||_2$ - 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 зануляются.