Линейные методы классификации

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

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

I семестр 2015 г.

Table of Contents

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:

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

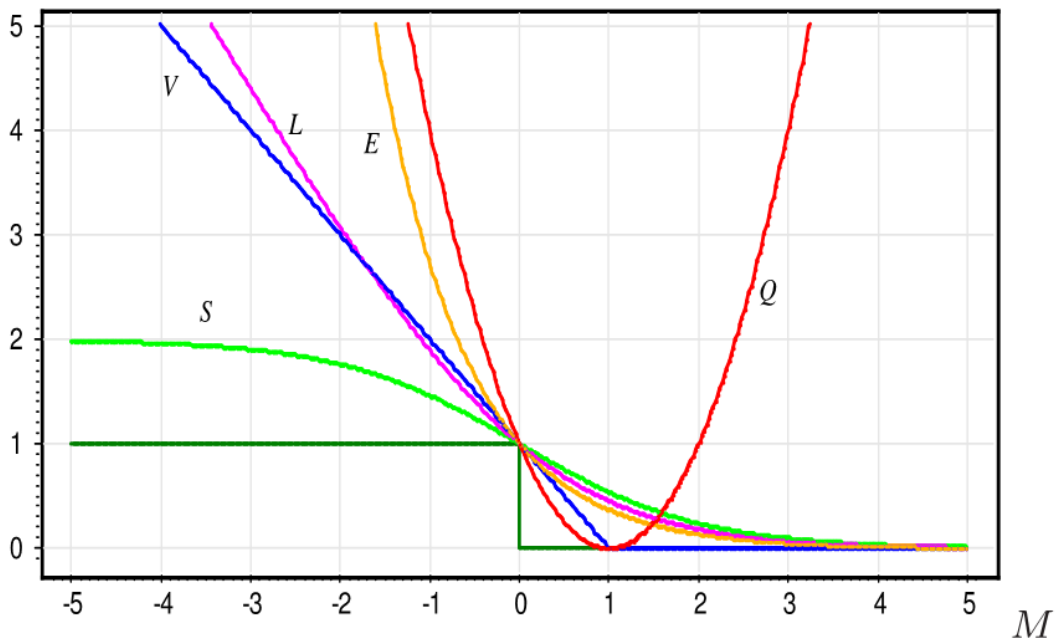
- Проблема: стандартные методы оптимизации неприменимы, т.к $Q(w, X)$ разрывна.
- Идея решения: аппроксимировать функцию цены сверху гладкой функцией $\mathcal{L}$:

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

Аппроксимация целевого критерия

Получаем аппроксимацию эмпирического риска сверху:

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



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

Table of Contents

Оптимизация

- Оптимизационная задача для получения весов:

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

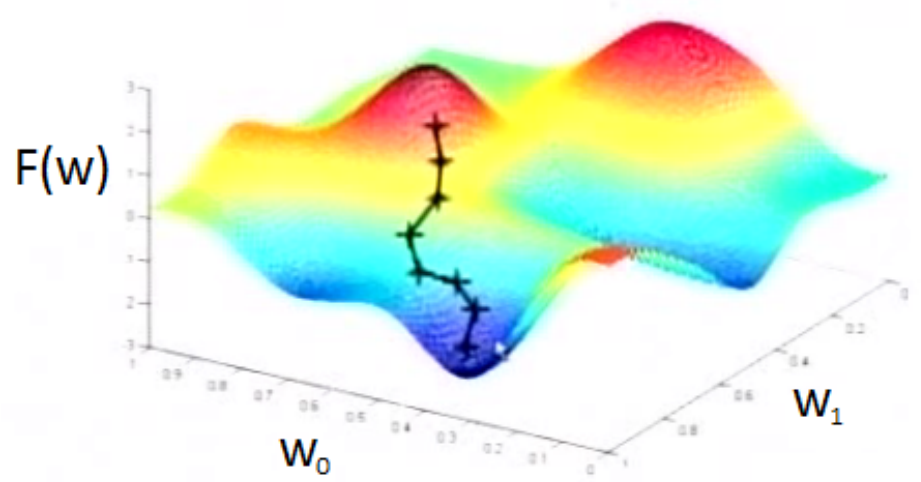
Алгоритм градиентного спуска

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

- $w_0 = w_1 = \ldots = w_D = 0$
- 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 $\mathcal{L}(u) = (u-1)^2$

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

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

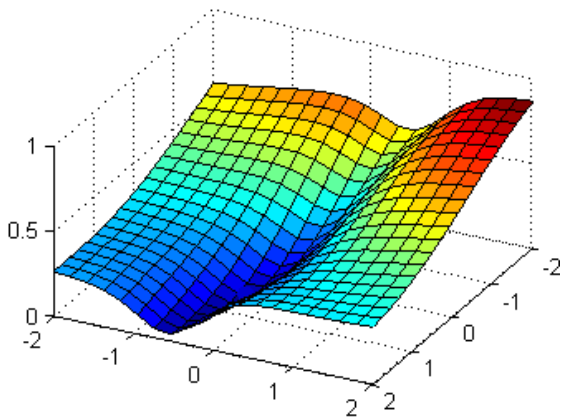
$\mathcal{L}(u) = [-u]_+$

Персептрон Розенблатта

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

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 (для $M < 0$ и отсутствии асимптот слева $\mathcal{L}'$ выше):
  - 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

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