(Квази)линейные методы классификации. Метод опорных векторов (Support Vector Machine – SVM)

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».
Лекция 3.

Общая характеристика

Метод обобщённого портрета (60–70-е годы, В.Н. Вапник и др.): отдалить объекты от разделяющей поверхности.
В 90-е годы метод стал называться машиной опорных векторов (support vector machine, SVM).

Свойства:

- сводится к эффективно решаемой задаче квадратичного программирования,
- разреженность (решение определяется опорными векторами),
- обобщается введением функции ядра.

Основные понятия

$X = R^n$ – пространство значений прогнозирующих переменных,
$Y = \{-1, 1\}$ – прогнозируемая переменная,
$D = X \times Y$.

Решающая функция (алгоритм классификации)

$$f : X \to Y.$$

$V = \{(x^i, y^i) \in D \,|\, i = \overline{1, N}\}$ – случайная независимая выборка, $V \in D^N$.

$Q\colon D^N \to \Phi$ — метод построения решающих функций,
Φ — заданный класс решающих функций.

Постановка задачи

Эмпирический риск:

$$\widetilde{R}(V, f) = \frac{1}{N} \sum_{i=1}^{N} I(y^i \neq f(x^i)),$$

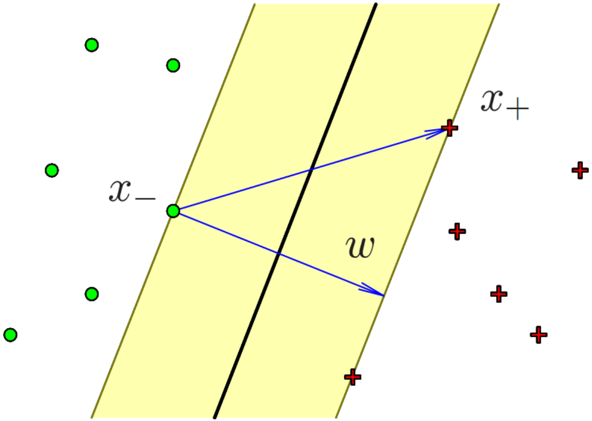
где $I(\cdot)$ – индикаторная функция.

Линейный пороговый классификатор

$$f(x) = \operatorname{sign}(wx - w_0).$$

Требуется найти вектор w и скаляр w_0, минимизирующие эмпирический риск при дополнительных требованиях.

Линейно разделимая выборка



Критерий

Нормировка

$$\min_{(x^i, y^i) \in V} y^i(x^i w - w_0) = 1.$$

Для граничных точек

$$x_+ w - w_0 = 1, \quad -(x_- w - w_0) = 1.$$

Ширина разделяющей полосы

$$(x_+ - x_-) \cdot \frac{w}{|w|} = \frac{(w_0 + 1) - (w_0 - 1)}{|w|} = \frac{2}{|w|}.$$

Оптимизационная задача

Задача квадратичной оптимизации

$$\begin{cases} w^2 \to \min\limits_{w,w_0} \\ y^i(x^i w - w_0) \geqslant 1, \quad i = 1, ..., N. \end{cases}$$

Условие нормировки выполняется автоматически.

Линейно неразделимая выборка

Задача квадратичного программирования

$$\begin{cases} \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \to \min\limits_{w, w_0, \xi} \\ y^i(x^i w - w_0) \geqslant 1 - \xi_i, \quad i = 1, ..., N, \\ \xi_i \geqslant 0, \quad i = 1, ..., N, \end{cases}$$

где $C > 0$ – параметр.

Применение метода

Оптимизационную задачу можно решать методом активных ограничений (incremental active set method, INCAS), частным случаем которого является симплекс-метод.

Константу C обычно выбирают по критерию скользящего контроля.

Задача сводится в задаче для линейно разделимой выборки, когда $C \to \infty$.

Функция Лагранжа

Задача на нахождение экстремума

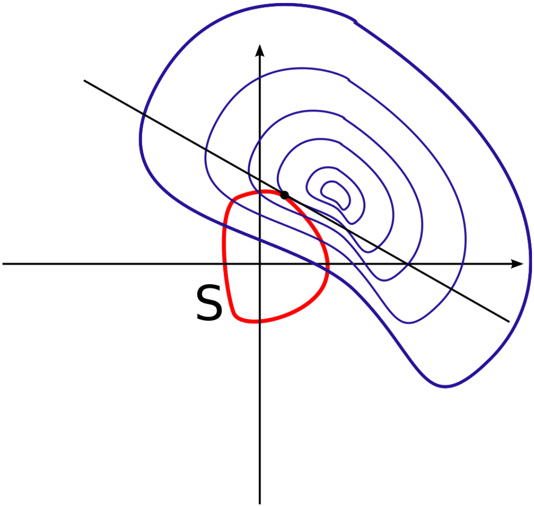
$$\begin{cases} f(x) \to \min\limits_x \\ \psi(x) = 0. \end{cases}$$

Условие касания линий уровня

$$\frac{\partial f(x)}{\partial x} = \lambda \frac{\partial \psi(x)}{\partial x}.$$

После переноса в левую часть получаем функцию Лагранжа.

Линии уровня и ограничение



Условия Каруша–Куна–Таккера

Задача на нахождение экстремума

$$\begin{cases} f(x) \to \min\limits_{x} \\ \psi(x) \geqslant 0. \end{cases}$$

Необходимые условия экстремума

$$\begin{cases} \frac{\partial f(x)}{\partial x} = \lambda \frac{\partial \psi(x)}{\partial x}, \\ \lambda \psi(x) = 0, \\ \lambda \geqslant 0. \end{cases}$$

Условия дополняющей нежёсткости и неотрицательности.

Функция Лагранжа для SVM

Функция Лагранжа

$$\mathfrak{L}(w, w_0, \xi; \lambda, \eta) = \frac{w^2}{2} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\lambda_i(y^i(x^i w - w_0) - 1 + \xi_i) - \sum_{i=1}^{N}\xi_i\eta_i.$$

После преобразований

$$\mathfrak{L}(w, w_0, \xi; \lambda, \eta) = \frac{w^2}{2} - \sum_{i=1}^{N}\lambda_i(y^i(x^i w - w_0) - 1) - \sum_{i=1}^{N}\xi_i(\lambda_i + \eta_i - C).$$

Условия стационарности

Дифферецируем

$$\begin{cases} \frac{\partial \mathfrak{L}}{\partial w} = 0, & w = \sum_{i=1}^{N} \lambda_i y^i x^i, \\ \frac{\partial \mathfrak{L}}{\partial w_0} = 0, & \sum_{i=1}^{N} \lambda_i y^i = 0, \\ \frac{\partial \mathfrak{L}}{\partial \xi_i} = 0, & \eta_i + \lambda_i = C, \quad i = 1, ..., N. \end{cases}$$

Двойственная задача

Выразим всё через λ_i

$$\begin{cases} -\mathfrak{L}(\lambda) = \sum\limits_{i=1}^{N} \lambda_i + \frac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} \lambda_i \lambda_j y^i y^j x^i x^j \to \min\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, ..., N, \\ \sum\limits_{i=1}^{N} \lambda_i y^i = 0. \end{cases}$$

Точки, для которых $\lambda_i > 0$, называются опорными векторами. Заметим, x^i что входят только через скалярные произведения.

Итоговый класификатор имеет вид

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N} \lambda_i y^i x^i x - w_0 \right).$$

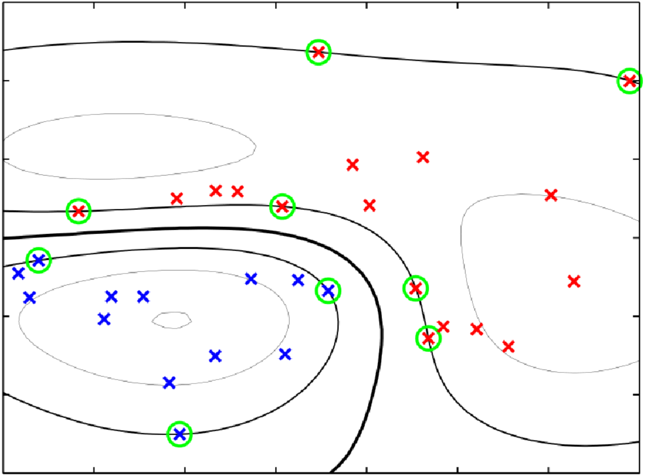
Ядра и спрямляющие пространства

Можно перейти от исходного пространства X к новому пространству H с помощью некоторого преобразования $\psi : X \to H$. Пространство H называют спрямляющим. После этого скалярное произведение примет вид

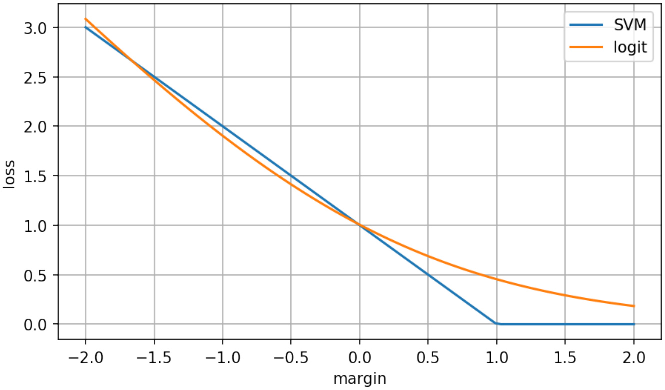
$$K(x, x') = \psi(x) \cdot \psi(x').$$

Можно вообще не вводить новое пространство, а сразу задать ядро K.

Пример работы SVM с ядром



SVM и логистическая регрессия



Оба метода эквивалентны безусловной минимизации соответствующей функции потерь с регуляризацией.

Открытые проблемы

- Как соотносится SVM с метрическими методами.
- Возможно ли придумать обобщение SVM, включающее дискриминант Фишера как частный случай.
- Является ли разреженность достоинством.
- Как выбирать ядро.
- Как применить kernel trick в логистической регрессии.
- Низкая эффективность при сильно перекрывающихся распределениях классов.

Общность линейных методов

- Дискриминант Фишера не требует вероятностных предположений, но по результату почти совпадает с линейным дискриминантом, основанным на нормальности.
- Наивный байесовский классификатор является частным случаем логистической регрессии.
- Метод SVM оказывается близок логистической регрессии ввиду сходства функций потерь.