Линейная классификация. Метод опорных векторов. ROC- кривая.

Московский физико-технический институт, МФТИ

Москва

Линейная модель классификации

Дано: выборка обучающих пар обьектов $X^l = (x_i, y_i)_{i=1}^l$. $X = \mathbb{R}^n$, $Y = \{-1, +1\}$. В общем виде алгоритм классификации представим функцией $a(x; w, w_0) = \text{sign}(\langle x, w \rangle)$. Рассмотрим классифицирующие модели вида $a(x, w) = \text{sign}(\langle x, w \rangle - w_0)$, так что множество значений функционала $Y = \{-1, +1\}$.

Функция доли неправильных ответов

$$\sum_{i=1}^{l} [a(x_i; w, w_0) \neq y_i] = \sum_{i=1}^{l} [M_i(w, w_0) \neq 0] \to \min_{w, w_0},$$

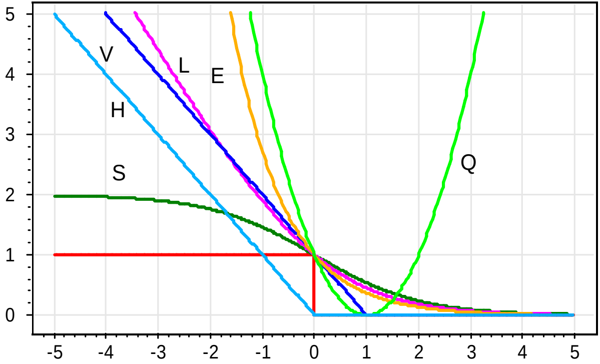
где $M_i(w, w_0) = (\langle w, x_i \rangle - w_0) y_i$ - отступ обьекта x_i.

Эмпирический риск - кусочно постоянная функция. Заменим оценкцой сверху, непрерывной по параметрам:

$$\sum_{i=1}^{l}[M_i(w, w_0) \neq 0] \leq \sum_{i=1}^{l}(1 - M_i(w, w_0))_+ + \frac{1}{2C}||w||^2 \to \min_{w,w_0}.$$

Аппроксимация - позволяет оценить близость обьектов к границе классов.
Регуляризация - регуяризация решения в случае мультиколлинеарности.

Аппроксимация пороговой функции



Пусть выборка $X^l = (x_i, y_i)_{i=1}^l$ линейно разделима:

$$\exists w, w_0 : \; M_i(w, w_0) = y_i(\langle w, x_i \rangle - w_0), \; i = 1, .., l.$$

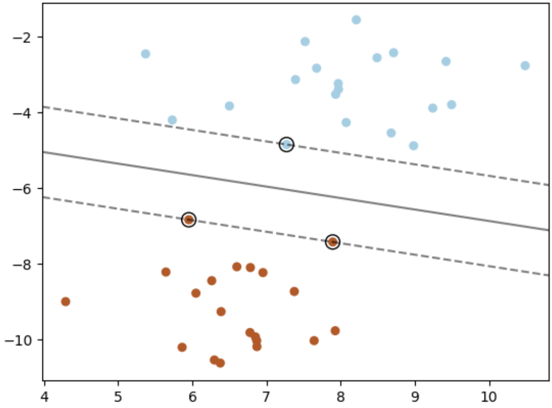
Отнормируем вектор $\min_{i=1,..,l} M_i(w, w_0) = 1$
Раздедяющая полоса

$$\begin{aligned} &\{x : -1 \le \langle w, x_i \rangle - w_0 \le 1\}, \\ &\exists x_- : \langle w, x_- \rangle - w_0 = -1, \\ &\exists x_+ : \langle w, x_+ \rangle - w_0 = 1. \end{aligned}$$

Ширина полосы

$$\frac{\langle x_+ - x_-, w \rangle}{||w||} = \frac{2}{||w||} \to \max.$$

SVM иллюстрация



Определим расстояние от произвольной точки $x_0 \in \mathbb{R}^d$

$$\rho(x_0, a) = \frac{|\langle w, x\rangle - w_0|}{||w||},$$

расстояние от гиперплоскости до ближайшего обьекта обучающей выборки равно

$$\min_{x \in X^l} \frac{|\langle w, x\rangle - w_0|}{||w||} = \frac{1}{||w||} \min_{x \in X} |\langle w, x\rangle - w_0| = \frac{1}{||w||}.$$

Разделимая выборка

Существует прямая (гиперплоскость) которая является линейным разграничителем для двух классов (Hard margin support vector machin):

$$\begin{cases} \frac{1}{2}||w||^2 \to \min\limits_{w, w_0}, \\ y_i(\langle w, x \rangle - w_0) \geq 1, \; i = 1, ..., l. \end{cases}$$

На практике линейно разделимые выборки встречаются достаточно редко. Постановку необхоимо модифицировать, так чтобы система ограничений была совместна в любом случае.

Неразделимая выборка

Хотя бы одно из ограничений в задаче нарушается

$$\exists x_i \in X : \; y_i(\langle w, x \rangle - w_0) < 1.$$

Введем штрафы

$$\forall x_i \in X : \; y_i(\langle w, x \rangle - w_0) \geq 1 - \xi_i, \; i = 1, .., l.$$

Задача оптимизации для неразделимой выборки

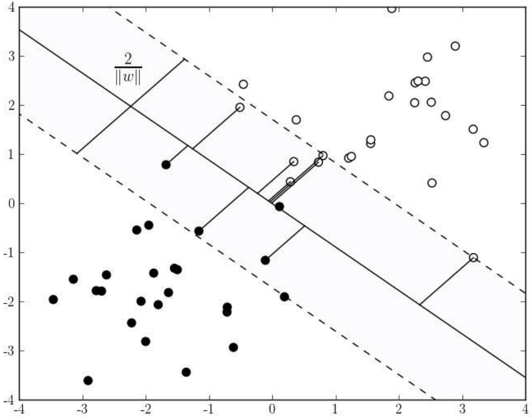
Задача максимизации отступа и минимизации штрафа противоположны одна другой. Оптимизационная задача для неразделимой выборки (soft margin support vector machine):

$$\begin{cases} \frac{1}{2}||w||^2 + C\sum\limits_{i=1}^{l}\xi_i \to \min\limits_{w,w_0} \\ M_i(w, w_0) = y_i(\langle w, x\rangle - w_0) \geq 1 - \xi_i,\ i = 1, ..., l \\ \xi \geq 0,\ i = 1, .., l. \end{cases}$$

Эквивалентна задаче безусловной минимизации:

$$C\sum_{i=1}^{l}(1 - M_i(w, w_0))_+ + ||w||^2 \to \min_{w,w_0}$$

SVM иллюстрация



Условия Каруша-Кунна-Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_{x}, \\ g_i(x) \leq 0,\ j = 1, .., l, \\ h_j(x) = 0,\ j = 1, .., k. \end{cases}$$

Необходимые условия. Если x - точка локального минимума, то существуют множители μ_i, $i = 1, .., m$, λ_j, $j = 1, .., k$.

Функционал $\mathcal{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{i=1}^{k} \lambda_j h_j(x)$,

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, \\ g_i(x) \leq 0,\ h_j(x) = 0, \\ \mu_j \geq 0, \\ \mu_i g_i(x) = 0. \end{cases}$$

Применение условий ККТ к задачам SVM

$$\mathcal{L}(w, w_0, \xi; \lambda, \eta) =$$

$$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}(1 - M_i(w, w_0))_+ - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C),$$

λ_i - переменная двойственная к ограничениям $M_i \geq 1 - \xi_i$;
η_i - переменная двойственная к ограничениям $\xi_i \geq 0$:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0, \ \frac{\partial \mathcal{L}}{\partial w_0} = 0, \ \frac{\partial \mathcal{L}}{\partial \xi} = 0, \\ \xi_i \geq 0, \ \lambda_i \geq 0, \ \eta_i \geq 0, \ i = 1, .., l; \\ \lambda_i = 0 \text{ либо } M_i(w, w_0) = 1 - \xi_i, \ i = 1, .., l, \\ \eta_i = 0 \text{ либо } \xi_i = 0, \ i = 1, .., l. \end{cases}$$

Необходимые условия седловой точки функций Лагранжа

$$\mathcal{L}(w, w_0, \xi; \lambda, \eta) =$$
$$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}(1 - M_i(w, w_0))_+ - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C), \quad (1)$$

Необходимые условия седловой точки функции Лагранжа:

$$\frac{\partial \mathbb{L}}{\partial w} = w - \sum_{i=1}^{l}\lambda_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{l}\lambda_i y_i x_i,$$
$$\frac{\partial \mathbb{L}}{\partial w_0} = -\sum_{i=1}^{l}\lambda_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{l}\lambda_i y_i,$$
$$\frac{\partial \mathbb{L}}{\partial \xi_i} = -\lambda_i - \eta_i + C \quad \Rightarrow \quad \lambda_i + \eta_i = C,\ i = 1, .., l.$$

Понятие опорного обькта

Рассмотрим различные возможные случаи решения задачи:

1. $\lambda_i = 0,\ \eta_i = C, \xi_i = 0,\ M_i \geq 1,$ - переферийные обьекты;
2. $0 < \lambda_i < C,\ 0 < \eta_i < C, \xi_i = 0,\ M_i = 1,$ - **опорные** граничные обьекты;
3. $\lambda_i = C,\ \eta_i = C, \xi_i > 0,\ M_i > 1,$ - **опорные** нарушители.

Решение чувствительно к выбросам.

Конечная задача

В итоге имеем задачу на минимизацию выпуклого фунционала на выпуклом множестве

$$\begin{cases} -\mathcal{L}(\lambda) = -\sum\limits_{i=1}^{l}\lambda_i + \frac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle \to \min\limits_{\lambda}, \\ 0 \le \lambda_i \le C,\ i = 1, .., l, \\ \sum\limits_{i=1}^{l}\lambda_i y_i = 0 \end{cases}$$

решение исходной задачи выражается через решение вспомогательной (двойственной)

$$\begin{cases} w = \sum\limits_{i=1}^{l}\lambda_i y_i x_i; \\ w_0 = \langle w, x_i\rangle - y_i, \forall i : \lambda_i > 0,\ M_i = 1 \end{cases}$$

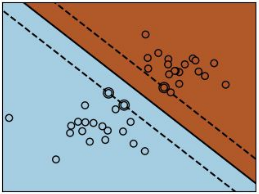
Итоговый классификатор представим в виде

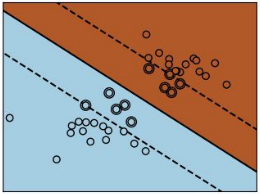
$$a(x) = \operatorname{sign}\left(\sum^{l}\lambda_i y_i\langle x_i, x\rangle - w_0\right).$$

Выбор ширины окна в зависимости от константы C

Рассмотрим задачу минимизации без ограничений

$$\sum_{i=1}^{l}(1 - M_i(w, w_0))_+ + \frac{1}{2C}||w||^2 \to \min_{w, w_0}$$





Нелинейное обобщение SVM

Определение

Фунция $K : X \times X \to \mathbb{R}$ *- ядро, если* $K(x, x') = \langle \psi(x), \psi'(x) \rangle$ *при некотором* $\psi : X \to H$*, где* H *- гилбертово пространство.*

Теорема

Функция $K(x, x')$ *является ядром тогда и только тогда, когда она симметрична* $K(x, x') = K(x', x)$ *: и неотрицательно определена*

$$\iint_{X \times X} K(x, x') g(x) g(x') dx dx' \geq 0 \, \text{для любого} \, g : X \to \mathbb{R}$$

.

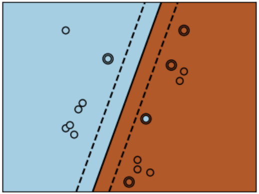
Правила построения ядерных функций

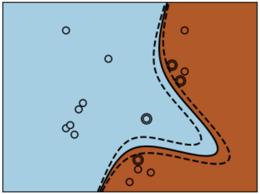
Примеры правил для построения различных ядер.

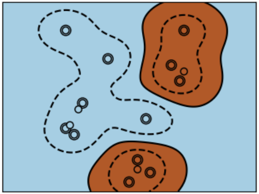
- $K(x, x') = \langle x, x' \rangle$,
- суперпозиция $K(x, x') = \alpha_1 K_1(x, x') + \alpha_2 K_2(x, x')$ $\alpha_1, \alpha_2 > 0$,
- произведение функций $K(x, x') = \psi(x)\psi(x')$,
- приозведение ядер $K(x, x') = K_1(x, x')K_2(x, x')$.

Классификация с различными ядрами

- линейное $\langle x, x' \rangle$;
- полиномиальное $(\langle x, x' \rangle + 1)^d$;
- гауссовское (RBF) $\exp(-\gamma ||x - x'||^2)$







Преимущества и недостатки SVM

Преимущества

- задача квадратичного программирования корректна поставлена, для нее разработаны эффектинвые методы решения
- сокращение размерности задачи до набора опорных векторов,
- максимизация зазора - аналог регуляризации.

Недостатки

- неустойчивость к шуму, алгоритм в качестве опорных берет обьекты -выбросы,
- нет универсального алгоритма построения ядер,
- подбор параметра регуляризации C затратная по времени задача.

Точность и полнота выборки

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN},$$

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN},$$

Критерий качества на основе точности и полноты: F - мера, гармоническое среднее точности и полноты:

$$F = \frac{2 * precision * recall}{precision + recall}$$

Площадь под ROC -кривой

Доли неверно принятых обьектов (False Positive Rate) и верно принятых обьектов (True Positive Rate)

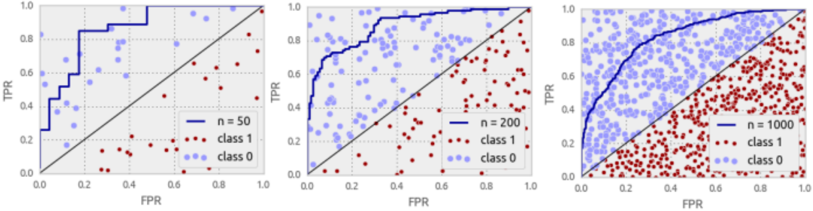
$$FPR = \frac{FP}{FP + TN},$$

$$FPR = \frac{TP}{TP + FN}.$$

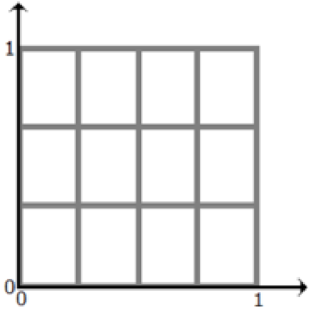
Проблема чувствительности к соотношению классов -> решение через precision-recall кривая оценки.

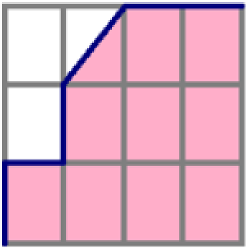
Иллюстрация

Пример ROC (receiver operating characteristic) кривой для различных наборов данных



Пример построения таблиц





Точность полученная в нашем случае AUC ROC = 9.5 / 12   0.79.

Пример классификации

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

id	> 0.25	класс
4	1	1
1	1	0
6	1	1
3	0	0
5	0	1
2	0	0
7	0	0

Табл. 3

Расличные возможные виды ROC кривой

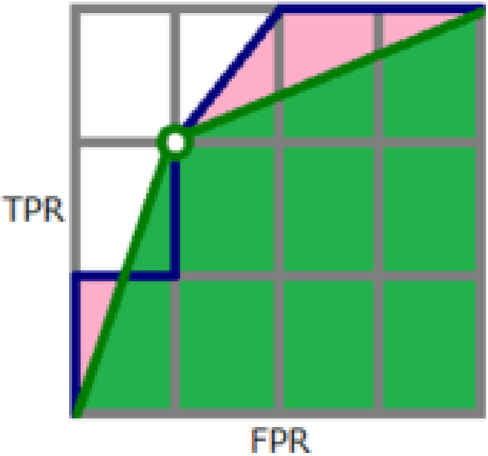
AUC ROC равен доле пар обьекетов вида (обьект класса 1, обьект класса 0), которые алгоритм верно упорядочил.

$$\frac{\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l} I[y_i < y_j] I'[a_i < a_j]}{\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l} I[y_i < y_j]},$$

$$I'[a_i < a_j] = \begin{cases} 0, & a_i > a_j, \\ 0.5 & a_i = a_j, \\ 1 & a_i < a_j, \end{cases}$$

Принятие решений

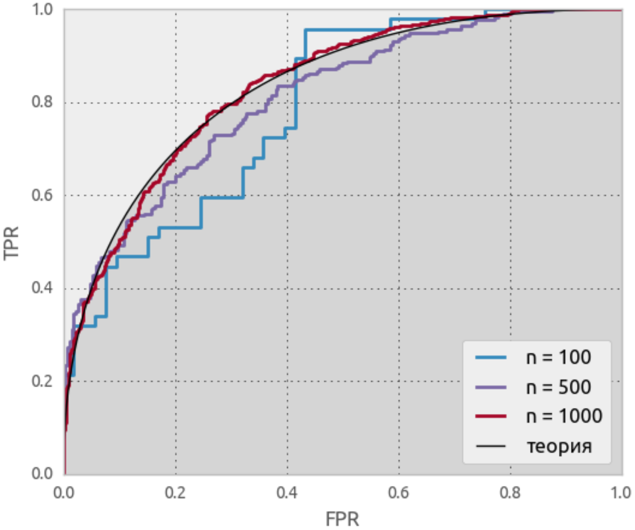
Необходимо выбрать порог значения, который определит пренадлежность классам 0 и 1.



точка (1/4,2/3) - соответствует критерию порога 1/4. Итак, в нашем случае, FPR -1/4, TPR - 2/3

Пример ROC -кривой

Введем распределени ебьктов.



Максимизация AUC ROC

Оптимизация непосрдественно затруднительна:

- функция недифференцируем по параметра алгоритма,
- в явном виде она не разбивается на отдельные слагаемые, которые зависят от ответа только на одном обьекте
- замена индикаторной функци на дифференцируемую,
- переход к выборке, состоящей из пар обектов,