Машина опорных векторов — является одной из наиболее популярных методологий обучения по прецедентам, предложенной В. Н. Вапником и известной в англоязычной литературе под названием SVM (Support Vector Machine).

Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin). Случай линейной разделимости. Задача квадратичного программирования. Опорные векторы. Случай отсутствия линейной разделимости. Функции ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Способы построения ядер. Примеры ядер. Сопоставление SVM и нейронной RBF-сети. Обучение SVM методом активных ограничений. SVM-регрессия.

Машина опорных векторов в задачах классификации

Понятие оптимальной разделяющей гиперплоскости

Линейный классификатор — алгоритм классификации, основанный на построении линейной разделяющей поверхности. В случае двух классов разделяющей поверхностью является гиперплоскость, которая делит пространство признаков на два полупространства. В случае большего числа классов разделяющая поверхность кусочно-линейна.

Определение

Пусть объекты описываются *n* числовыми признаками $f_j\colon X \to \mathbb{R},\ j = 1,\ldots,n$. Тогда пространство признаковых описаний объектов есть $X = \mathbb{R}^n$. Пусть Y — конечное множество номеров (имён, меток) классов.

Случай двух классов

Положим $Y = \{-1,+1\}$.

Линейным классификатором называется алгоритм классификации $a\colon X \to Y$ вида

$$a(x,w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = \operatorname{sign}\langle x,w\rangle,$$

где w_j — вес j-го признака, w_0 — порог принятия решения, $w = (w_0,w_1,\ldots,w_n)$ — вектор весов, $\langle x,w\rangle$ — скалярное произведение признакового описания объекта на вектор весов.

Предполагается, что искусственно введён «константный» нулевой признак: $f_0(x) = -1$.

Случай произвольного числа классов

Линейный классификатор определяется выражением

$$a(x,w) = \arg\max_{y \in Y} \sum_{j=0}^{n} w_{yj} f_j(x) = \arg\max_{y \in Y} \langle x, w_y \rangle,$$

где каждому классу соотвествует свой вектор весов $w_y = (w_{y0}, w_{y1}, \ldots, w_{yn})$

Обучение линейного классификатора

Метод минимизации эмпирического риска

Обучение (настройка) линейного классификатора методом минимизации эмпирического риска заключается в том, чтобы по заданной обучающей выборке пар «объект, ответ» $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ построить алгоритм a: $X \to Y$ указанного вида, минимизирующий *фунционал эмпирического риска*:

$$Q(w) = \sum_{i=1}^{m} [a(x_i, w) \neq y_i] \to \min_w.$$

Методы обучения линейных классификаторов различаются подходами к решению данной оптимизационной задачи.

Понятие отступа

В случае двух классов, $Y = \{-1, +1\}$, удобно определить для произвольного обучающего объекта $x_i \in X^m$ величину *отступа* (margin):

$$M(x_i) = y_i \langle x_i, w \rangle.$$

В случае произвольного числа классов отступ определяется выражением

$$M(x_i) = \langle x_i, w_{y_i} \rangle - \max_{y \in Y,\ y \neq y_i} \langle x_i, w_y \rangle.$$

Отступ можно понимать как «степень погруженности» объекта в свой класс. Чем меньше значение отступа $M(x_i)$, тем ближе объект подходит к границе классов, тем выше становится вероятность ошибки. Отступ $M(x_i)$ отрицателен тогда и только тогда, когда алгоритм $a(x)$ допускает ошибку на

объекте x_i. Это наблюдение позволяет записать фунционал эмпирического риска в следующем виде:

$$Q(w) = \sum_{i=1}^{m} [M(x_i) < 0].$$

Замена пороговой функции потерь

Минимизация функционала $Q(w)$ по вектору весов сводится к поиску максимальной совместной подсистемы в системе неравенств. Эта задача является NP-полной и может иметь очень много решений, поскольку минимальное число ошибок может реализоваться на различных подмножествах объектов. Однако абсолютно точное решение этой задачи, и, тем более, нахождение всех её решений, в большинстве приложений не представляет практического интереса. Обычно вполне устраивает приближённое решение, достаточно близкое к точному.

Наиболее известные методы обучения линейного классификатора связаны с заменой пороговой функции потерь её различными непрерывными аппроксимациями:

$$[M < 0] \leq L(M),$$

где $L: \mathbb{R} \to \mathbb{R}_+$ — непрерывная или гладкая функция, как правило, невозрастающая.

После замены функции потерь минимизируется не сам функционал эмпирического риска, а его верхняя оценка.

$$Q(w) \leq \widetilde{Q}(w) = \sum_{i=1}^{m} L(M(x_i)).$$

Применение аппроксимаций имеет ряд преимуществ.

- Некоторые аппроксимации способны улучшать обобщающую способность классификатора. В частности, известно, что пробит-аппроксимация при некоторых условиях уменьшает вероятность ошибки [Langford, McAllester].
- Непрерывные аппроксимации позволяют применять известные численные методы оптимизации для настройки весов w_j, в частности, градиентные методы и методы выпуклого программирования.

Регуляризация

Наряду с заменой пороговой функции потерь, рекомендуется добавлять к функционалу штрафное слагаемое, запрещающее слишком большие значения нормы вектора весов:

$$Q(w) \leq \widetilde{Q}(w) = \sum_{i=1}^{m} L(M(x_i)) + \gamma \|w\|^p \to \min_w.$$

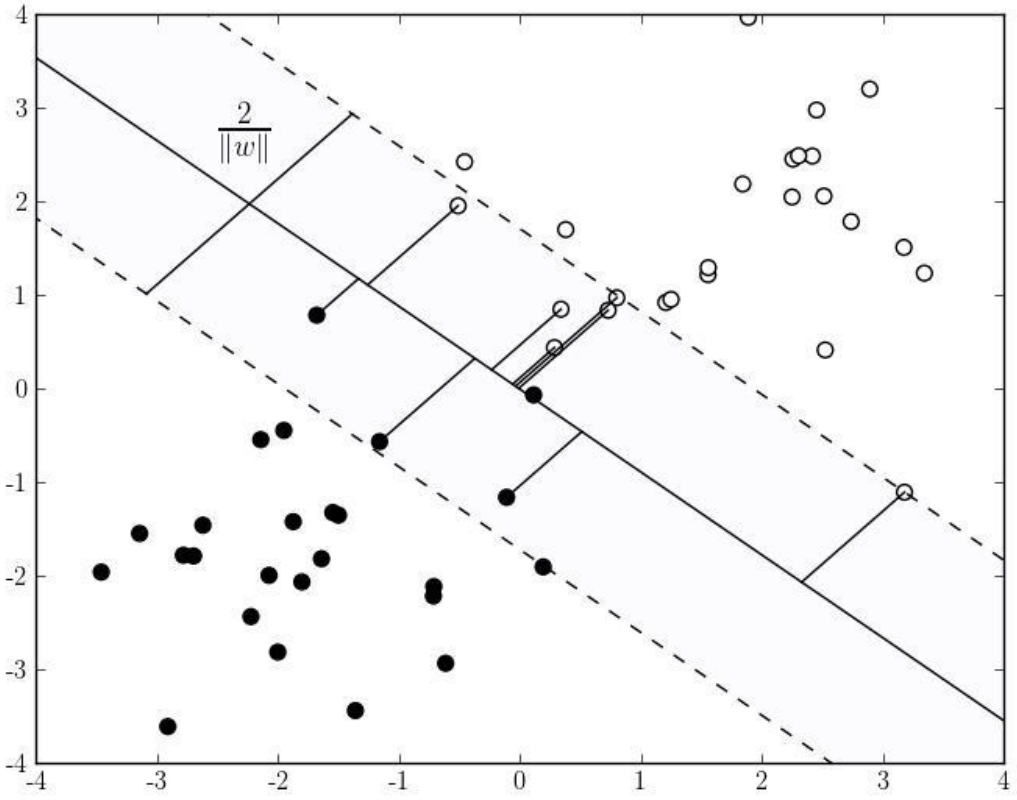
Добавление штрафного слагаемого или регуляризация снижает риск переобучения и повышает устойчивость вектора весов по отношению к малым изменениям обучающей выборки. Идея регуляризации предлагалась разными авторами, в разные годы, для разных алгоритмов, и называлась, соотвественно, по разному: сокращение весов (weght decay) — в нейронных сетях, гребневая регрессия (ridge regression) — в регрессионном анализе, сжатие весов (shrinkage).

Параметр регуляризации γ подбирается исходя из априорных соображений, либо по скользящему контролю. Существуют также теоретические оценки обобщающей способности линейного классификатора, позволяющие оценивать параметр регуляризации по явным формулам.

Степень регуляризатора p определяет класс методов оптимизации.

- При $p = 2$ и гладком (по параметру w) функционале $Q(w)$ можно применять стандартные градиентные методы минимизации.
- При $p = 1$ и выпуклом функционале $Q(w)$ возникает задача выпуклого программирования с ограничениями типа неравенств. В результате её решения часть коэффициентов w_j обнуляются, что фактически означает отсев неинформативных признаков.

Линейно разделимая выборка



Рассмотрим задачу нахождения наилучшего в некотором смысле разделения множества векторов на два класса с помощью линейной решающей функции. Пусть имеется множество прецедентов (Ξ, Y), где $\Xi = \{x_1,\dots,x_N\}$ — обучающая выборка, а $Y = (y_1,\dots,y_N)$ — множество меток двух классов ω_1 и ω_2. Требуется по обучающей выборке построить линейную решающую функции, т.е. такую линейную функцию $f(x)$, которая удовлетворяла бы условию

$f(x_i) > 0$ для всех $x_i \in \omega_1$,

$f(x_i) < 0$ для всех $x_i \in \omega_2$.

Без ограничения общности можно считать, что метки классов равны

$$y_i = \begin{cases} 1, & x_i \in \varpi_1, \\ -1, & x_i \in \varpi_2. \end{cases}$$

Тогда поставленную выше задачу можно переформулировать следующим образом. Требуется найти линейную решающую функцию $f(x)$, которая бы удовлетворяла условию

$y_i f(x_i) > 0$ для всех $x_i \in \Xi$ (1)

Умножая, если нужно, функцию f на некоторое положительное число, нетрудно видеть, что система неравенств (1) равносильна системе

$y_i f(x_i) > 1$ для всех $x_i \in \Xi$.

Кроме того, так как $f(x)$ — линейная функция, то последняя система неравенств примет вид

$y_i((w,x_i)+b) \geq 1, \; i = 1,...,N$, (2)

где w — вектор весовых коэффициентов, b — некоторое число. Тогда разделяющей два класса гиперплоскостью будет $(w,x)+b=0$. Нетрудно видеть, что и все гиперплоскости вида $(w,x)+b'=0$, где $b' \in (b-1,b+1)$, также будут разделяющими (рис.1). Расстояние между граничными гиперплоскостями $(w,x)+b-1=0$ и $(w,x)+b+1=0$ равно $\frac{2}{||w||}$.
Действительно, и — нормальные уравнения этих гиперплоскостей.
Тогда $p_1 = \frac{b-1}{||w||}$ и $p_2 = \frac{b+1}{||w||}$ — расстояния от этих гиперплоскостей до начала координат и $\frac{2}{||w||}$ — расстояние между гиперплоскостями. На самих граничных плоскостях может находиться некоторое число обучающих векторов. Эти векторы называются *опорными*.

Для надежного разделения классов необходимо, чтобы расстояние между разделяющими гиперплоскостями было как можно большим, т.е. $||w||$была как можно меньше. Таким образом, ставится задача нахождения минимума квадратичного функционала $0.5(w,w)$ (коэффициент 0.5 вводится для удобства дифференцирования) в выпуклом многограннике, задаваемым системой неравенств (2). В выпуклом множестве квадратичный функционал всегда имеет единственный минимум (если это множество не пусто). Из теоремы Куна — Таккера следует, что решение этой оптимизационной задачи равносильно поиску седловой точки лагранжиана

$$L(w,b,\lambda) = 0.5(w,w) - \sum_{i=1}^{N} \lambda_i (y_i((w,x_i)+b)-1) \rightarrow \min_{w,b} \max_{\lambda}$$

в ортанте по множителям Лагранжа $\lambda_i \geq 0 \quad (i = 1,...,N)$, при условии, что

$$\lambda_i\left(y_i((w,x_i)+b)-1\right)=0,\ i=1,...,N.$$

Последнее условие равносильно тому, что

$$\lambda_i=0 \text{ или } y_i((w,x_i)+b)-1=0,\ i=1,...,N \qquad (3)$$

Из необходимых условий существования седловой точки (полагая $x_i=(x_{i1},x_{i2},...,x_{in})$) имеем

$$\begin{cases} 0=\frac{\partial L}{\partial w_j}=w_j-\sum_{i=1}^{N}\lambda_i y_i x_{ij},\ j=1,...,n, \\ 0=\frac{\partial L}{\partial b}=\sum_{i=1}^{N}\lambda_i y_i. \end{cases}$$

Откуда следует, что вектор w следует искать в виде

$$w=\sum_{i=1}^{N}\lambda_i y_i x_i, \qquad (4)$$

причем

$$\sum_{i=1}^{N}\lambda_i y_i=0. \qquad (5)$$

В силу (3) в сумму (4) с ненулевыми коэффициентами λ_i входят только те векторы, для которых $y_i((w,x_i)+b)-1=0$. Такие векторы называют опорными, так как это именно те векторы, через которые будут проходить граничные гиперплоскости, разделяющие классы. Для найденного весового вектора w смещение b можно вычислить как $b=y_s{}^{-1}-(w,x_s)$ для любого опорного вектора x_s.

Найдем значения множителей Лагранжа, как критических точек лагранжиана. Для этого подставим (4) и (5) в лагранжиан, получим

$$L(w,b,\lambda)=0.5(w,w)-\sum_{i=1}^{N}\lambda_i(y_i((w,x_i)+b)-1)=$$

$$0.5(w,w)-\left((w,w)-\sum_{i=1}^{N}\lambda_i\right)=\sum_{i=1}^{N}\lambda_i-0.5(w,w)=$$

$$\sum_{i=1}^{N}\lambda_i - 0.5\sum_{i,j=1}^{N}\lambda_i\lambda_j y_i y_j (x_i, x_j) = \sum_{i=1}^{N}\lambda_i - 0.5\left\|\sum_{i=1}^{N}\lambda_i y_i x_i\right\|^2.$$

Таким образом, задача сводится к нахождению критических точек функции

$$\Phi(\lambda) = \sum_{i=1}^{N}\lambda_i - 0.5\left\|\sum_{i=1}^{N}\lambda_i y_i x_i\right\|^2 \qquad (6)$$

Так как эта функция представляет собой разность линейной и квадратичной функций, причем квадратичная функция отрицательно определена, то требуется найти наибольшее значение функции $\Phi(\lambda)$ при условии $\sum_{i=1}^{N}\lambda_i y_i = 0$ в области $\lambda_i \geq 0 \quad (i = 1,...,N)$. Существует много алгоритмов (в теории оптимизации) решения этой задачи (например, градиентные методы, метод покоординатного спуска и т.д.).

Замечания.

1. Суммирования в (6) осуществляются не по всем векторам, а только по опорным, которых может быть гораздо меньше, чем обучающих.
2. Линейная решающая функция в результате имеет вид , где λ_i зависят только y_i и от значений скалярного произведения (x_i, x_j), причем суммирования осуществляются только по опорным векторам.
3. После того, как решающая функция $f(x)$ вычислена, вектор x следует относить классу ω_1, если $f(x) > 0$ и классу ω_2, если $f(x) < 0$. Вероятность неправильной классификации можно оценить с помощью некоторой непрерывно убывающей функции $\varphi(t)$, удовлетворяющей условиям: $\varphi(0) = 0.5$, $\varphi(t) \to 0$ при $t \to \infty$. Тогда вероятность $p(x)$ неправильной классификации вектора x будет равна $\varphi(\rho(x, L_i))$, если $x \in \omega_i$ $(i = 1,2)$, где $L_i: (w,x) + b + \mathrm{sgn}(\alpha - i) = 0$, $1 < \alpha < 2$. То есть $p(x) = \varphi(\left|\left(\frac{\omega}{\|w\|}, x\right) + \frac{b + sgn(\alpha - i)}{\|w\|}\right|)$, если $x \in \omega_i$ $(i = 1,2)$.
4. В такой постановке алгоритм линейный классификации был разработан В. Вапником в 1963 году.

Пример. Методом опорных векторов разделите классы $\omega_1=\{x_1\}$ $p(x)=\varphi\left(\left|\left(\frac{w}{||w||},x\right)+\frac{b+\mathrm{sgn}(\alpha-i)}{||w||}\right|\right)$ и $\omega_2=\{x_2,x_3\}$, если $x_1=(1,1)^T$, $x_2=(1,2)^T$, $x_3=(2,3)^T$.

Решение. Положим $y_1=1$, $y_2=-1$, $y_3=-1$. Тогда функция $\Phi(\lambda)$ будет иметь вид

$$\Phi(\lambda)=\sum_{i=1}^{3}\lambda_i-0.5\sum_{i,j=1}^{3}\lambda_i\lambda_j y_i y_j(x_i,x_j)=$$
$$\lambda_1+\lambda_2+\lambda_3-0.5(2\lambda_1^2+5\lambda_2^2+13\lambda_3^2-6\lambda_1\lambda_2-10\lambda_1\lambda_3+16\lambda_2\lambda_3),$$

причем $\lambda_1-\lambda_2-\lambda_3=0\Rightarrow\lambda_3=\lambda_1-\lambda_2$. Тогда $\Phi(\lambda_1,\lambda_2)=2\lambda_1-2.5\lambda_1^2-\lambda_2^2+3\lambda_1\lambda_2$. Составим и решим нормальную систему для функции $\Phi(\lambda_1,\lambda_2)$:

$$\begin{cases}\frac{\partial\Phi}{\partial\lambda_1}=0,\\ \frac{\partial\Phi}{\partial\lambda_2}=0\end{cases}\Leftrightarrow\begin{cases}2-5\lambda_1+3\lambda_2=0,\\ -2\lambda_2+\lambda_1=0\end{cases}\Leftrightarrow\begin{cases}\lambda_1=4,\\ \lambda_2=6.\end{cases}$$

Следовательно, $\lambda_1=4$, $\lambda_2=6$, $\lambda_3=-2$. Так как $\lambda_3<0$, то исследуем функцию $\Phi(\lambda)$ на границе области $\lambda_i\geq 0\quad(i=1,2,3)$ при условии $\lambda_3=\lambda_1-\lambda_2$. Если $\lambda_1=0$, то $\lambda_3=-\lambda_2\Rightarrow\lambda_i^{(1)}=0\quad(i=1,2,3)\Rightarrow\Phi(\lambda^{(1)})=0$. Пусть $\lambda_2=0$. Тогда $\lambda_1=\lambda_3=\lambda$ и $\Phi(\lambda)=2\lambda-2.5\lambda^2$, $\Phi'(\lambda)=0$ при $\lambda^{(2)}=2/5$. Следовательно, $\lambda_1^{(2)}=\lambda_3^{(2)}=2/5$, $\lambda_2^{(2)}=0$ и $\Phi(\lambda^{(2)})=2/5$.

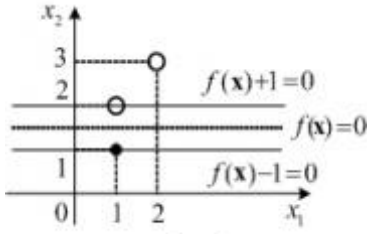


Рис.2

Если же $\lambda_3=0$, то $\lambda_1=\lambda_2=\lambda$ и $\Phi(\lambda)=2\lambda-0.5\lambda^2$, $\Phi'(\lambda)=0$ при $\lambda^{(3)}=2$. Следовательно, $\lambda_1^{(3)}=\lambda_2^{(3)}=2$, $\lambda_3^{(3)}=0$ и $\Phi(\lambda^{(3)})=2$.

Таким образом, наибольшее значение функции $\Phi(\lambda)$ в области $\lambda_i\geq 0$ ($i=1,2,3$) при условии $\lambda_3=\lambda_1-\lambda_2$ достигается в точке $\lambda^{(3)}=(2,2,0)^T$. В этом случае,

$$\begin{cases} w = \sum_{i=1}^{3} \lambda_i y_i x_i = 2x_1 - 2x_2 = 2\binom{1}{1} - 2\binom{1}{2} = \binom{0}{-2}, \\ b = \frac{1}{y_1} - (w, x_1) = 1 - (0 - 2) = 3. \end{cases}$$

Таким образом, $f(x) = (w,x) + b = -2x_2 + 3$ и $f(x) = 0 \Leftrightarrow x_2 = 1.5$. Ширина разделяющей полосы будет равна $\frac{2}{\|w\|}$, а прямые $f(x) + 1 = 0 \Leftrightarrow f(x) = 0 \Leftrightarrow x_2 = 2$ и $f(x) - 1 = 0 \Leftrightarrow x_2 = 1$ будут ее границами (см. рис.2).

Линейно неразделимая выборка

В 1992 году в работе Бернарда Бозера (Boser B.), Изабелл Гийон (Guyon I.) и Владимира Вапника был предложен способ адаптации машины опорных векторов для нелинейного разделения классов. В этом случае нужно вложить пространство признаков R^n в пространство H большей размерности с помощью отображения $\varphi: R^n \to H$. Будем считать, что H пространство со скалярным произведением. Тогда, рассматривая алгоритм опорных векторов для образов $\varphi(x_i)$ обучающей выборки, сведем решение задачи к линейно разделимому случаю, т.е. разделяющую функцию будем искать в виде

$$f(x) = \big(w, \varphi(x)\big) + b, \quad w = \sum_{i=1}^{N} \lambda_i y_i \varphi(x_i),$$

где коэффициенты λ_i зависят от y_i и от значения $\big(\varphi(x_i), \varphi(x_j)\big)$. Таким образом, для нахождения решающей функции нужно знать значения скалярных произведений $\big(\varphi(x_i), \varphi(x_j)\big)$. Для этого исследуем свойства функции $K(x,y) = \big(\varphi(x), \varphi(y)\big)$, которая называется ядром. Следующая теорема, известная в теории интегральных операторов и доказанная Джеймсом Мерсером в 1909 году, полностью характеризует ядро.

Теорема 1. Функция $K(x,y)$ является ядром тогда и только тогда, когда она удовлетворяет условиям:

1. $K(x,y) = K(y,x)$ (симметричность);
2. $K(x,y)$ неотрицательно определена, т.е. матрица $K = (K_{i,j})$, $K_{i,j} = K(x_i, x_j)$ является неотрицательно определенной для любых векторов $x_1, \ldots, x_m$.

Упражнение. Докажите, что следующие функции являются ядрами:

1. $K(x,y)=(x,y)$;
2. $K(x,y)=\varphi(x)\varphi(y)$;
3. $K(x,y)=C>0$.

Теорема 2. Справедливы следующие свойства ядер:

1. сумма ядер – ядро;
2. произведение ядер – ядро;
3. сумма равномерно сходящегося ряда ядер – ядро;
4. композиция ядра и любого отображения (т.е. $K(\psi(x),\psi(y))$) – ядро.

Следствие.

1. многочлен с положительными коэффициентами от ядра – ядро;
2. экспонента от ядра – ядро;
3. функция $e^{-||x-y||^2}$ ядро.

Доказательство. Утверждения 1 и 2 следуют из пунктов 1, 2 и 3 теоремы. Справедливость утверждения 3 вытекает из того, что $e^{-||x-y||^2}=e^{-(x_1-y_1)^2}\cdot...\cdot e^{-(x_n-y_n)^2}$, а симметричность и положительная определенность функций $e^{-(x_i-y_i)^2}$ проверяется непосредственно.

Любые $m+1$ векторов могут быть разделены на любые два класса с помощью мономиального отображения степени не больше m. Поэтому, если $\varphi: x \to \left\{x_1^{i_1}...x_n^{i_n}\right\}$, $i_1+...+i_n \le m$ такое отображение, то ядро, соответствующее этому отображению можно искать в виде

$$K(x,y)=\left(\varphi(x),\varphi(y)\right)=\left((x,y)+1\right)^m.$$

Таким образом, это ядро гарантирует разделение любых $m+1$ векторов на любы$e^{-||x-y||^2}=e^{-(x_1-y_1)^2}\cdot...\cdot e^{-(x_n-y_n)^2}$е два класса. В этом случае нахождение разделяющих функций осуществляется следующим образом:

1) найдем наибольшее значение функции

$$\Phi(\lambda)=\sum_{i=1}^{N}\lambda_i-0.5\sum_{i,j=1}^{N}\lambda_i\lambda_j y_i y_j K(x_i,x_j)$$

при условии $\sum_{i=1}^{N}\lambda_i y_i=0$ в области $\lambda_i\geq 0 \quad (i=1,...,N)$, получим вектор $\lambda^{(0)}=\left(\lambda_1^{(0)},...,\lambda_N^{(0)}\right)$;

2) разделяющую функцию ищем в виде

$$f(x)=(w,\varphi(x))+b=\left(\sum_{i=1}^{N}\lambda_i^0 y_i\varphi(x_i),\varphi(x)\right)+b=$$

$$\sum_{i=1}^{N}\lambda_i^0 y_i\left(\varphi(x_i),\varphi(x)\right)+y_r-1-\sum_{i=1}^{N}\lambda_i^0 y_i\left(\varphi(x_i),\varphi(x_r)\right)=$$

$$\sum_{i=1}^{N}\lambda_i^0 y_i K(x_i,x)+y_r-1-\sum_{i=1}^{N}\lambda_i^0 y_i K(x_i,x_r).$$

Преимущества и недостатки SVM:

- это наиболее быстрый метод нахождения решающих функций;
- метод сводится к решению задачи квадратичного программирования в выпуклой области, которая всегда имеет единственное решение;
- метод находит разделяющую полосу максимальной ширины, что позволяет в дальнейшем осуществлять более уверенную классификацию;
- метод чувствителен к шумам и стандартизации данных;
- не существует общего подхода к автоматическому выбору ядра (и построению спрямляющего подпространства в целом) в случае линейной неразделимости классов.

Больше примеров можно найти по ссылке
http://www.machinelearning.ru/wiki/index.php?title=SVM

Источник: http://www.machinelearning.ru