支持向量机 SVM 分类

支持向量机(support vector machine,SVM)是一种新的机器学习方法,其基础是 Vapnik 创建的统计学习理论(statistical learning theory,STL)。统计学习理论采用结构风险最小化(structural risk minimization,SRM)准则,在最小化样本点误差的同时,最小化结构风险,提高了模型的泛化能力,且没有数据维数的限制。在进行线性分类时,将分类面取在离两类样本距离较大的地方;进行非线性分类时通过高维空间变换,将非线性分类变成高维空间的线性分类问题。

一般支持向量机分类采用的是线性可分 SVM，以下为模型概述。

支持向量机最初是研究线性可分问题而提出的,因此,这里先详细介绍线性 SVM 的基本思想及原理。

不失一般性,假设大小为 l 的训练样本集$\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,l\}$由两个类别组成,若 $\boldsymbol{x}_i$ 属于第一类,则记 $y_i=1$;若 $\boldsymbol{x}_i$ 属于第二类,则记 $y_i=-1$。

若存在分类超平面

$$\boldsymbol{w}\boldsymbol{x}+\boldsymbol{b}=0 \tag{28-1}$$

能够将样本正确地划分成两类,即相同类别的样本都落在分类超平面的同一侧,则称该样本集是线性可分的。即满足

$$\begin{cases}\boldsymbol{w}\boldsymbol{x}_i+\boldsymbol{b}\geqslant 1, & y_i=1\\ \boldsymbol{w}\boldsymbol{x}_i+\boldsymbol{b}\leqslant -1, & y_i=-1\end{cases},\qquad i=1,2,\cdots,l \tag{28-2}$$

定义样本点 $\boldsymbol{x}_i$ 到式(28-1)所指的分类超平面的间隔为

$$\varepsilon_i=y_i(\boldsymbol{w}\boldsymbol{x}_i+b)=|\boldsymbol{w}\boldsymbol{x}_i+b| \tag{28-3}$$

将式(28-3)中的 $\boldsymbol{w}$ 和 $\boldsymbol{b}$ 进行归一化,即用$\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$和$\frac{\boldsymbol{b}}{\|\boldsymbol{w}\|}$分别代替原来的 $\boldsymbol{w}$ 和 b,并将归一化后的间隔定义为几何间隔

$$\delta_i=\frac{\boldsymbol{w}\boldsymbol{x}_i+\boldsymbol{b}}{\|\boldsymbol{w}\|} \tag{28-4}$$

同时,定义一个样本集到分类超平面的距离为此集合中与分类超平面最近的样本点的几何间隔,即

$$\delta=\min\delta_i,\qquad i=1,2,\cdots,l \tag{28-5}$$

样本的误分次数 N 与样本集到分类超平面的距离 δ 间的关系为

$$N \leqslant \left(\frac{2R}{\delta}\right)^2 \tag{28-6}$$

其中，$R=\max\|x_i\|$，$i=1,2,\cdots,l$，为样本集中向量长度最长的值。

由式(28-6)可知，误分次数 N 的上界由样本集到分类超平面的距离 δ 决定，即 δ 越大，N 越小。因此，需要在满足式(28-2)的无数个分类超平面中选择一个最优分类面，使得样本集到分类超平面的距离 δ 最大。

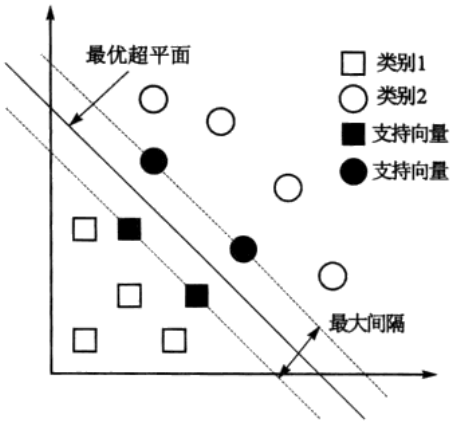


图 28-1　最优超平面示意图

若间隔 $\varepsilon=|\boldsymbol{w}\boldsymbol{x}_i+\boldsymbol{b}|=1$，则两类样本点间的距离为 $2\dfrac{|\boldsymbol{w}\boldsymbol{x}_i+\boldsymbol{b}|}{\|\boldsymbol{w}\|}=\dfrac{2}{\|\boldsymbol{w}\|}$。因此，如图 28-1 所示，目标即为在满足式(28-2)的约束下寻求最优分类超平面，使得 $\dfrac{2}{\|\boldsymbol{w}\|}$ 最大，即最小化 $\dfrac{\|\boldsymbol{w}\|^2}{2}$。用数学语言描述，即

$$\begin{cases}\min \dfrac{\|\boldsymbol{w}\|^2}{2} \\ \text{s.t. } y_i(\boldsymbol{w}\boldsymbol{x}_i+\boldsymbol{b})\geqslant 1, \qquad i=1,2,\cdots,l\end{cases} \tag{28-7}$$

该问题可以通过求解 Largrange 函数的鞍点得到，即

$$\Phi(\boldsymbol{w},\boldsymbol{b},\alpha_i)=\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{l}\alpha_i[y_i(\boldsymbol{w}\boldsymbol{x}_i+b)-1] \tag{28-8}$$

其中，$\alpha_i>0$，$i=1,2,\cdots,l$，为 Largrange 系数。

由于计算的复杂性，一般不直接求解，而是依据 Largrange 对偶理论将式(28-8)转化为对偶问题，即

$$\begin{cases}\max Q(\boldsymbol{\alpha})=\sum\limits_{i=1}^{l}\alpha_i-\dfrac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\boldsymbol{x}_j) \\ \text{s.t. } \sum\limits_{i=1}^{l}\alpha_i y_i=0, \qquad \alpha_i\geqslant 0\end{cases} \tag{28-9}$$

这个问题可以用二次规划方法求解，设求解得到的最优解为 $\boldsymbol{\alpha}^*=[\alpha_1^*,\alpha_2^*,\cdots,\alpha_l^*]^{\mathrm{T}}$，则可以得到最优的 $\boldsymbol{w}^*$ 和 b^* 为

$$\begin{cases}\boldsymbol{w}^*=\sum\limits_{i=1}^{l}\alpha_i^*\boldsymbol{x}_i y_i \\ b^*=-\dfrac{1}{2}\boldsymbol{w}^*(\boldsymbol{x}_r+\boldsymbol{x}_s)\end{cases} \tag{28-10}$$

其中，$\boldsymbol{x}_r$ 和 $\boldsymbol{x}_s$ 为两个类别中任意的一对支持向量。

最终得到的最优分类函数是

$$f(\boldsymbol{x})=\operatorname{sgn}\left[\sum_{i=1}^{l}\alpha_i^* y_i(\boldsymbol{x}\boldsymbol{x}_i)+b^*\right] \tag{28-11}$$

值得一提的是，若数据集中的绝大多数样本是线性可分的，仅有少数几个样本(可能是异常点)导致寻找不到最优分类超平面。针对此类情况，通用的做法是引入松弛变量，并对式(28-7)

中的优化目标及约束项进行修正，即

$$\begin{cases} \min \dfrac{\|w\|^2}{2} + C\sum_{i=1}^{l}\xi_i \\ \text{s.t.} \begin{cases} y_i(wx_i + b) \geqslant 1 - \xi_i \\ \xi_i > 0 \end{cases}, \quad i = 1,2,\cdots,l \end{cases} \tag{28-12}$$

其中，C 为惩罚因子，起着控制错分样本惩罚程度的作用，从而实现在错分样本的比例与算法复杂度间的折中。求解方法与式(28－8)相同，即转化为其对偶问题，只是约束条件变为

$$\begin{cases} \sum_{i=1}^{l}\alpha_i y_i = 0 \\ 0 \leqslant \alpha_i \leqslant C \end{cases}, \quad i = 1,2,\cdots,l \tag{28-13}$$

最终求得的分类函数的形式与式(28－11)一样。

支持向量机分类的数学原理

设样本集为 $\{(x_i, y_i) \mid x_i \in R^n; y_i \in \{-1, +1\}, i = 1,...,I\}$，我们的目的是寻找一个最优超平面 H 使得标签为＋1 和－1 的两类点不仅分开且分得间隔最大。

当在 n 维欧几里德空间中就可以实现线性分离时，也即存在超平面将样本集按照标签－1 与＋1 分在两边。由于超平面在 n 维欧几里德空间中的数学表达式是一个线性方程 $< w, x > + b = 0$，其中，w 为系数向量，x 为 n 维变量，$< w, x >$ 内积，b 为常数。空间中点 x_i 到超平面 L 的距离 $d(x_i, L) = \dfrac{|< w, x_i > + b|}{\| w \|}$。欲使得 $d(x_i, H)$ 最大，等价于 $\dfrac{1}{2}\| w \|^2$ 最小。于是，得到一个在约束条件下的极值问题

$$\begin{cases} \min \dfrac{1}{2}\| w \|^2 \\ y_i(< w, x_i > + b) \geq 1, \quad i = 1,2,...,I \end{cases}$$

引入 Lagrange 乘子 $\alpha = (\alpha_1, \alpha_2, ..., \alpha_I)$，可以解得关于该参变量的方程

$$Q(\alpha) = \sum_{i=1}^{I}\alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{I}\alpha_i\alpha_j y_i y_j < x_i, x_j >$$

称之为 Lagrange 对偶函数。其约束条件为

$$\sum_{i,j=1}^{I}\alpha_i y_i = 0, \ \alpha_i \geq 0, i = 1,2,..., I$$

在此约束条件之下， 使得 $Q(\alpha)$ 达到最大值的 α 的许多分量为 0，不为 0 的 α_i 所对应的样本 x_i 就称为支持向量。这就是支持向量的来历。

当在输入空间不能实现线性分离，假设我们找到了非线性映射 ϕ 将样本集 $\{(x_i, y_i) \mid x_i \in R^n; y_i \in \{-1, +1\}, i = 1,...,I\}$ 映射到高维特征空间 H 中，此时我们考虑在 H 中的集合 $\{(\phi(x_i), y_i) \mid x_i \in R^n; y_i \in \{-1, +1\}, i = 1,...,I\}$ 的线性分类，即在 H 中构造超平面，其权系数 w 满足类似的极值问题。由于允许部分点可以例外，那么可以引入松弛项，即改写为：

$$\begin{cases} \min \dfrac{1}{2}\| w \|^2 + C\sum_{i=1}^{L}\xi_i \\ y_i(< w, x_i > + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ i = 1,2,...,I \end{cases}$$

最终转化为一个二次型在约束条件下的二次规划问题：

$$\begin{cases}\min_{\alpha}\dfrac{1}{2}\alpha' D\alpha + c'\alpha \\ y'\alpha = 0,\ 0 \le \alpha = (\alpha_1,...,\alpha_I)^T \le A = (C,...,C)^T\end{cases}$$

其中，$y=(y_1,...,y_I)^T$，$c=(-1,...,-1)^T$，$D=\left(K(x_i,x_j)y_iy_j\right)_{1\le i,j\le I}$ 为矩阵。$K(x,s)$ 是核函数。

一分类问题是一个极端情形但却又是非常有用的，它可以表示为如下数学模型：设 $\left\{x_i \mid x_i \in R^n, i=1,...,I\right\}$ 为空间 R^n 的有限观测点，找一个以 a 为心，以 R 为半径的包含这些点的最小球体。因此，一分类是对于求一个化合物成分的最小包络曲面的最佳方法。与前面完全相同的手法，设 ϕ 是由某个核函数 $K(x,s)$ 导出的从输入空间到特征空间中的嵌入映射，最后可以得到二次规划问题

$$\begin{cases}\min_{\alpha}\dfrac{1}{2}\alpha' D\alpha + c'\alpha \\ y'\alpha = 0,\ 0 \le \alpha = (\alpha_1,...,\alpha_I)^T \le A = (C,...,C)^T\end{cases}$$

其中，$y=(y_1,...,y_I)^T$，$c=(-1,...,-1)^T$，$D=\left(K(x_i,x_j)y_iy_j\right)_{1\le i,j\le I}$ 为矩阵。$K(x,s)$ 是核函数。此时

$$f(x)=K(x,x)-2\sum_{i=1}^{L}\alpha_iK(x,x_i)+\sum_{j=1}^{L}\sum_{i=1}^{L}\alpha_i\alpha_jK(x_i,x_j)$$

此时几乎所有的点满足 $f(x)\le R^2$ 。参数 C 起着控制落在球外点的数目,变化区间为：$1/L<C<1$.

案例①-线性可分 SVM

（程序已详解，模型步骤就不列出了）

序号	fenlei.txt								类别
1	8.35	23.53	7.51	8.62	17.42	10	1.04	11.21	1
2	9.25	23.75	6.61	9.19	17.77	10.48	1.72	10.51	1
3	8.19	30.5	4.72	9.78	16.28	7.6	2.52	10.32	1
4	7.73	29.2	5.42	9.43	19.29	8.49	2.52	10	1
5	9.42	27.93	8.2	8.14	16.17	9.42	1.55	9.76	1
6	9.16	27.98	9.01	9.32	15.99	9.1	1.82	11.35	1
7	10.06	28.64	10.52	10.05	16.18	8.39	1.96	10.81	1
8	9.09	28.12	7.4	9.62	17.26	11.12	2.49	12.65	1
9	9.41	28.2	5.77	10.8	16.36	11.56	1.53	12.17	1
10	8.7	28.12	7.21	10.53	19.45	13.3	1.66	11.96	1
11	6.93	29.85	4.54	9.49	16.62	10.65	1.88	13.61	1
12	8.67	36.05	7.31	7.75	16.67	11.68	2.38	12.88	1
13	9.98	37.69	7.01	8.94	16.15	11.08	0.83	11.67	1
14	6.77	38.69	6.01	8.82	14.79	11.44	1.74	13.23	1
15	8.14	37.75	9.61	8.49	13.15	9.76	1.28	11.28	1
16	7.67	35.71	8.04	8.31	15.13	7.76	1.41	13.25	1
17	7.9	39.77	8.49	12.94	19.27	11.05	2.04	13.29	1
18	7.18	40.91	7.32	8.94	17.6	12.75	1.14	14.8	1
19	8.82	33.7	7.59	10.98	18.82	14.73	1.78	10.1	1
20	6.25	35.02	4.72	6.28	10.03	7.15	1.93	10.39	1
21	10.6	52.41	7.7	9.98	12.53	11.7	2.31	14.69	2

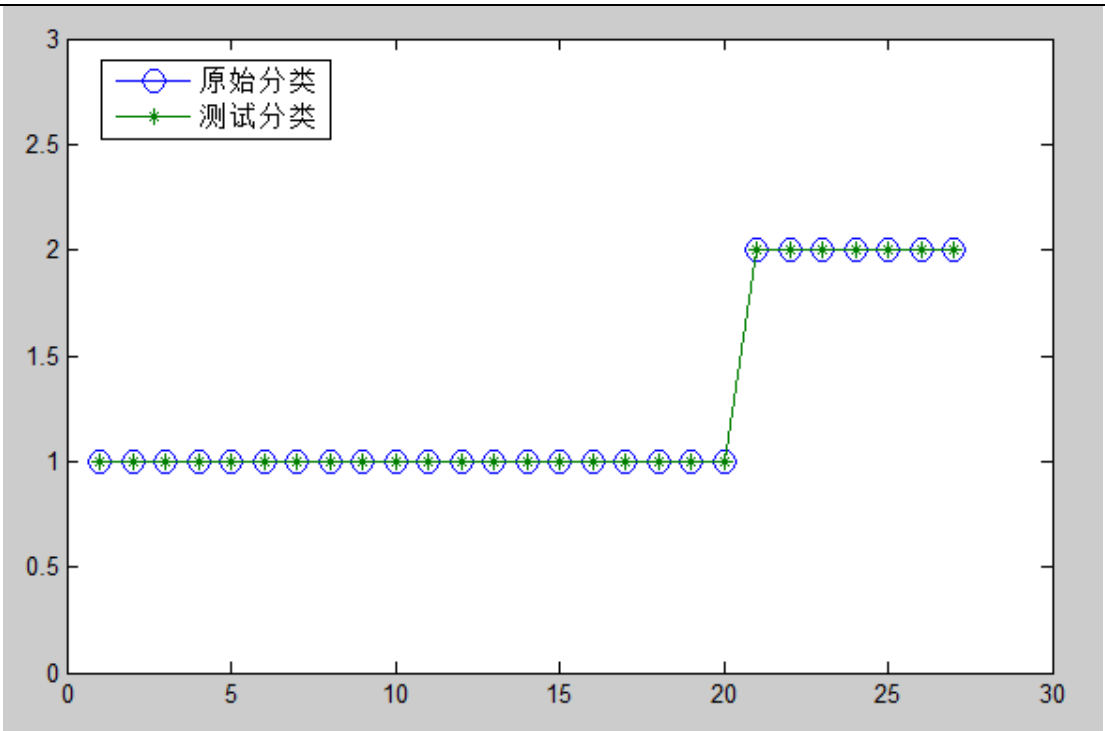
22	7.27	52.65	3.84	9.16	13.03	15.26	1.98	14.57	2
23	13.45	55.85	5.5	7.45	9.55	9.52	2.21	16.3	2
24	10.85	44.68	7.32	14.51	17.13	12.08	1.26	11.57	2
25	7.21	45.79	7.66	10.36	16.56	12.86	2.25	11.69	2
26	7.68	50.37	11.35	13.3	19.25	14.59	2.75	14.87	2
27	7.78	48.44	8	20.51	22.12	15.73	1.15	16.61	2
28	7.94	39.65	20.97	20.82	22.52	12.41	1.75	7.9	待分类
29	8.28	64.34	8	22.22	20.06	15.12	0.72	22.89	待分类
30	12.47	76.39	5.52	11.24	14.52	22	5.46	25.5	待分类

将蓝色数据录入文本文件中，本程序将文本文件命名为“fenlei”
需要 SVM 函数工具箱才能运行

附录 1	运行环境：Matlab2011a

```
clc, clear
a0=load('fenlei.txt'); %把表中 x1...x8 的所有数据保存在纯文本文件 fenlei.txt 中
a=a0'; b0=a(:,[1:27]); dd0=a(:,[28:end]); %提取已分类和待分类的数据
[b,ps]=mapstd(b0); %已分类数据的标准化
%mapstd 按行逐行地对数据进行标准化处理，
%将每一行数据分别标准化为均值为 ymean(默认为 0)、
%标准差为 ystd(默认为 1)的标准化数据，其计算公式是：y = (x-xmean)*(ystd/xstd) + ymean。
%如果设置的 ystd=0，或某行的数据全部相同(此时 xstd =0)，
%存在除数为 0 的情况，则 Matlab 内部将此变换变为 y = ymean。
dd=mapstd('apply',dd0,ps); %待分类数据的标准化
group=[ones(20,1); 2*ones(7,1)]; %已知样本点的类别标号，即设置分类，
%本程序设置前 20 个为第一类，21-27 为 2 类
s=svmtrain(b',group) %训练支持向量机分类器
sv_index=s.SupportVectorIndices   %返回支持向量的标号（分类器）
beta=s.Alpha   %返回分类函数的权系数（分类器）
bb=s.Bias   %返回分类函数的常数项（分类器）
mean_and_std_trans=s.ScaleData %第 1 行返回的是已知样本点均值向量的相反数，
%第 2 行返回的是标准差向量的倒数（分类器）
check=svmclassify(s,b')   %验证已知样本点
%将检验图画出，更直观些（可不画）
x=1:27;%样本数据有 27 个
a=group';
b=check';
axis([0,28,0,3]);%设置坐标轴范围
plot(x,a,'-o',x,b,'-*')
err_rate=1-sum(group==check)/length(group) %计算已知样本点的错判率
solution=svmclassify(s,dd') %对待判样本点进行分类
```

检验图



待分类 28，29，30 的分类情况为

```
solution =

     2
     2
     2
```

案例②-线性可分 SVM

该程序为恶性肿瘤数据分类，由于数据较多，大家在调试的时候请自行找数据，该案例程序与案例①相似

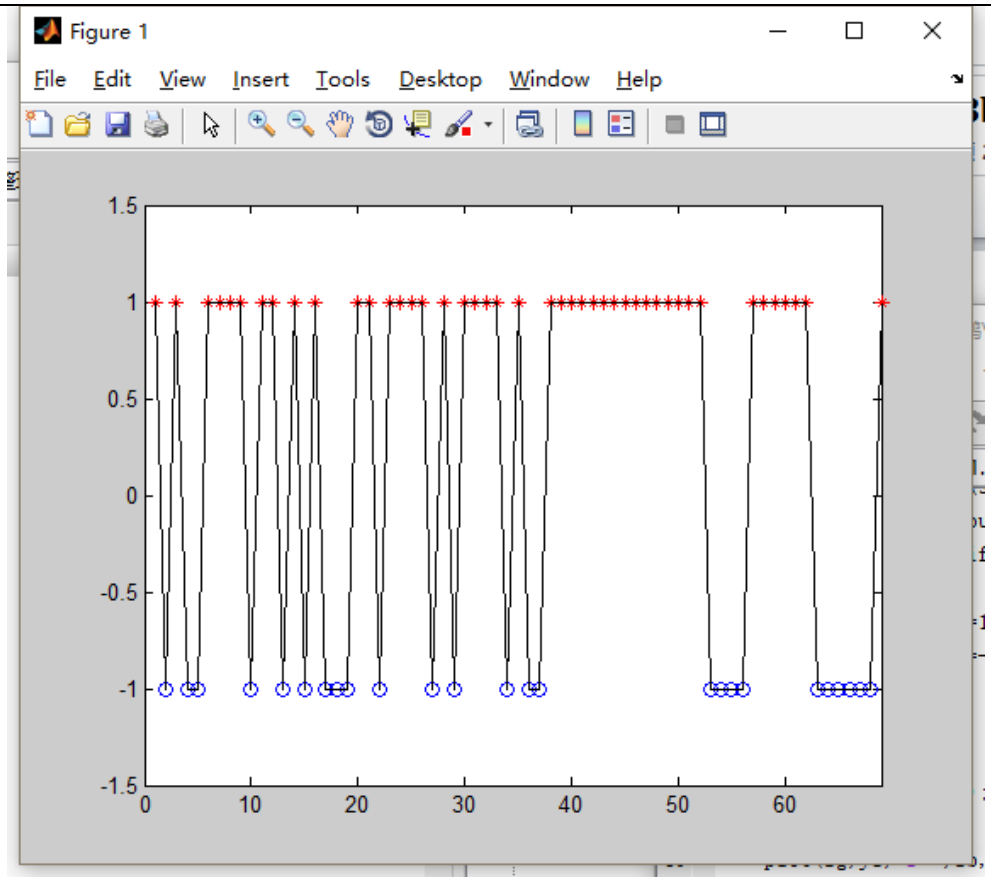
```
842302,-1,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053
842517,-1,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.
84300903,-1,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999,0.7456,0.786
84348301,-1,11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744,0.4956,1.
84358402,-1,20.29,14.34,135.1,1297,0.1003,0.1328,0.198,0.1043,0.1809,0.05883,0.7572,0.78
843786,-1,12.45,15.7,82.57,477.1,0.1278,0.17,0.1578,0.08089,0.2087,0.07613,0.3345,0.8902
844359,-1,18.25,19.98,119.6,1040,0.09463,0.109,0.1127,0.074,0.1794,0.05742,0.4467,0.7732
84458202,-1,13.71,20.83,90.2,577.9,0.1189,0.1645,0.09366,0.05985,0.2196,0.07451,0.5835,1
844981,-1,13,21.82,87.5,519.8,0.1273,0.1932,0.1859,0.09353,0.235,0.07389,0.3063,1.002,2.
84501001,-1,12.46,24.04,83.97,475.9,0.1186,0.2396,0.2273,0.08543,0.203,0.08243,0.2976,1.
845636,-1,16.02,23.24,102.7,797.8,0.08206,0.06669,0.03299,0.03323,0.1528,0.05697,0.3795,
84610002,-1,15.78,17.89,103.6,781,0.0971,0.1292,0.09954,0.06606,0.1842,0.06082,0.5058,0.
846226,-1,19.17,24.8,132.4,1123,0.0974,0.2458,0.2065,0.1118,0.2397,0.078,0.9555,3.568,11
846381,-1,15.85,23.95,103.7,782.7,0.08401,0.1002,0.09938,0.05364,0.1847,0.05338,0.4033,1
84667401,-1,13.73,22.61,93.6,578.3,0.1131,0.2293,0.2128,0.08025,0.2069,0.07682,0.2121,1.
84799002,-1,14.54,27.54,96.73,658.8,0.1139,0.1595,0.1639,0.07364,0.2303,0.07077,0.37,1.0
848406,-1,14.68,20.13,94.74,684.5,0.09867,0.072,0.07395,0.05259,0.1586,0.05922,0.4727,1.
84862001,-1,16.13,20.68,108.1,798.8,0.117,0.2022,0.1722,0.1028,0.2164,0.07356,0.5692,1.0
```

将以上数据录入文本文件中，本程序将文本文件命名为“cancerdata”，该数据第一列为编号，不用管，第二列是 1 和-1，指的是良性和恶性，后面的就是各指标数据了。

附录 2	运行环境：Matlab2011a

```
clc,clear
a=load('cancerdata.txt');
a(:,1)=[];   %删除第一列病例号
gind=find(a(:,1)==1);   %读出良性肿瘤的序号
bind=find(a(:,1)==-1); %读出恶性肿瘤的序号
training=a([1:500],[2:end]); %提出已知样本点的数据
training=training';
[train,ps]=mapstd(training); %已分类数据标准化
group(gind)=1; group(bind)=-1;   %已知样本点的类别标号
group=group'; %转换成列向量
xa0=a([501:569],[2:end]); %提出待分类数据
xa=xa0'; xa=mapstd('apply',xa,ps); %待分类数据标准化
s=svmtrain(train',group, 'Method','SMO', 'Kernel_Function','quadratic') %使用序列最小化方法训练支持向量机的分类器，
%如果使用二次规划的方法训练支持向量机则无法求解
sv_index=s.SupportVectorIndices'   %返回支持向量的标号
beta=s.Alpha'   %返回分类函数的权系数
b=s.Bias   %返回分类函数的常数项
mean_and_std_trans=s.ScaleData %第 1 行返回的是已知样本点均值向量的相反数，
%第 2 行返回的是标准差向量的倒数
check=svmclassify(s,train');   %验证已知样本点
err_rate=1-sum(group==check)/length(group) %计算错判率
solution=svmclassify(s,xa'); %进行待判样本点分类
solution=solution'
sg=find(solution==1)   %求待判样本点中的良性编号
sb=find(solution==-1) %求待判样本点中的恶性编号
%画图
y1=1;
y2=-1;
n=length(solution);
x=1:n;
plot(sg,y1,'r*',sb,y2,'bo',x,solution,'k-')
axis([0 n -1.5 1.5])
```

运行结果



结果在矩阵 solution 中