计算视觉与模式识别

分类器

数据集

$$(x_1, y_1) \quad \cdots \quad (x_n, y_n)$$

分类

$$x_{\text{new}} \rightarrow \ ?$$

风险函数

损失函数

$$L(i \to j) = \begin{cases} l > 0 & i \neq j \\ 0 & i = j \end{cases}$$

风险函数

$$R(s) = Pr\{1 \to 2 | \text{using } s\} L(1 \to 2) + Pr\{2 \to 1 | \text{using } s\} L(2 \to 1)$$

分类边界

$$
\begin{aligned}
& P\{\text{class is } 2|x\}L(2\to 1)+P\{\text{class is } 1|x\}L(1\to 1) \\
= \; & P\{\text{class is } 2|x\}L(2\to 1)+0 \\
= \; & p(2|x)L(2\to 1) \\
& P\{\text{class is } 1|x\}L(1\to 2) \\
= \; & p(1|x)L(1\to 2)
\end{aligned}
$$

边界上

$$p(2|x)L(2\to 1) = p(1|x)L(1\to 2)$$

得

$$p(x|2)p(2)L(2\to 1) = p(x|1)p(1)L(1\to 2)$$

分类

$$
\begin{array}{lll}
x \to 1 & p(x|2)p(2)L(2 \to 1) & < \; p(x|1)p(1)L(1 \to 2) \\
x \to 2 & p(x|2)p(2)L(2 \to 1) & > \; p(x|1)p(1)L(1 \to 2)
\end{array}
$$

多类别贝叶斯分类器

损失函数

- $\exists k, \Pr(k|x) > 1 - d$

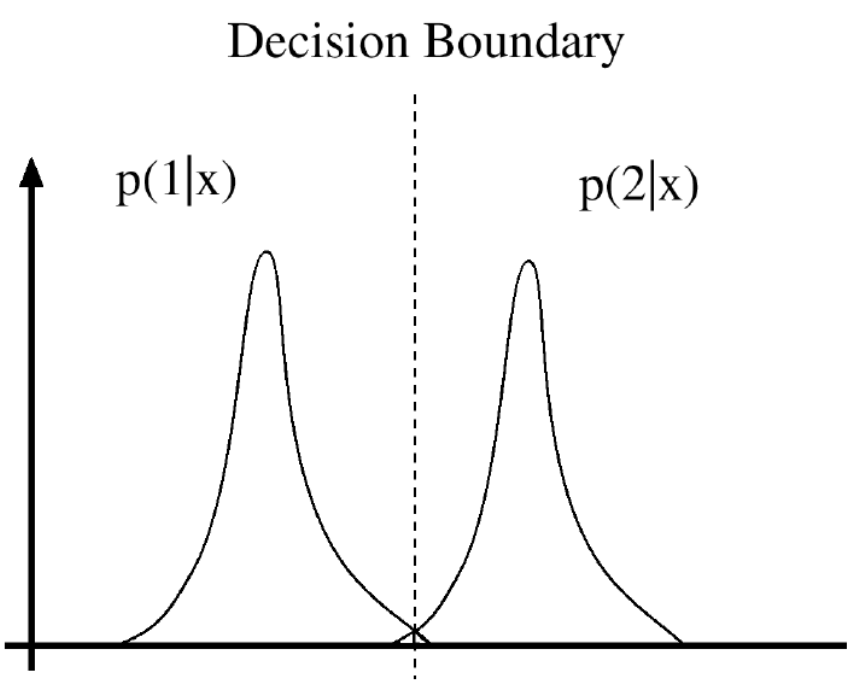
$$L(i \to j) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

- $\forall k, \Pr(k|x) < 1 - d$,d<1

$$L(i \to j) = d$$

选择类别

$$c = \arg\max_k \Pr(k|x)$$



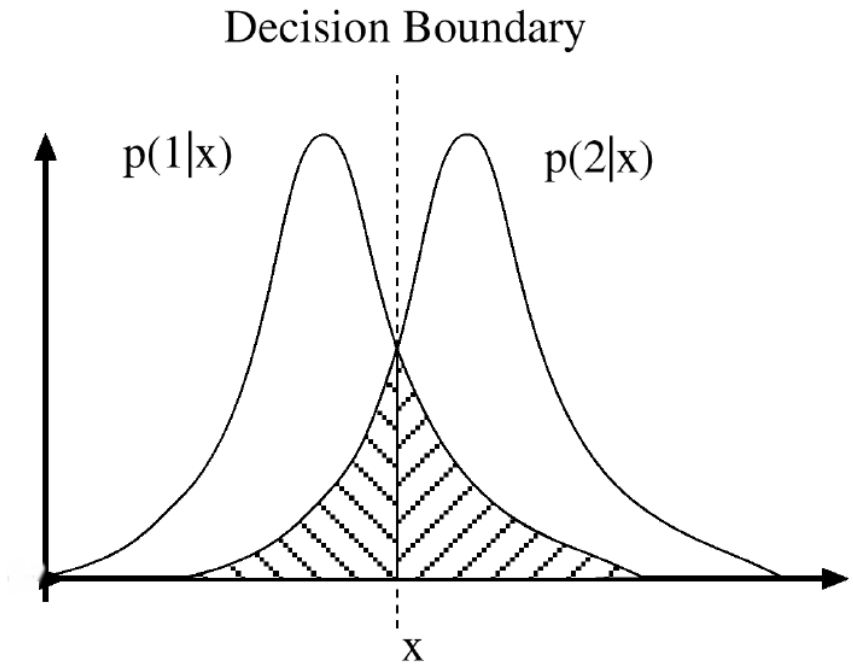


Figure 25.1. This figure shows typical elements of a two class classification problem. We have plotted $p(\text{class}|x)$ as a function of the feature x. Assuming that $L(1 \to 2) = L(2 \to 1)$, we have marked the classifier boundaries. In this case, the Bayes risk is the sum of the amount of the posterior for class one in the class two region and the amount of the posterior for class two in the class one region (the hatched area in the figures). For the case on the left, the classes are well separated, which means that the Bayes risk will be small; for the case on the right, the Bayes risk is rather large.

Logistic regression

$$\log\frac{p(1|\boldsymbol{x})}{p(-1|\boldsymbol{x})} = \boldsymbol{a}^T\boldsymbol{x}$$

$$p(1|\boldsymbol{x}) = \frac{e^{\boldsymbol{a}^T\boldsymbol{x}}}{1+e^{\boldsymbol{a}^T\boldsymbol{x}}}$$

$$p(-1|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{a}^T\boldsymbol{x}}}$$

$$L = -\sum_i \frac{1+y_i}{2}\boldsymbol{a}^T\boldsymbol{x}_i - \ln(1+e^{\boldsymbol{a}^T\boldsymbol{x}_i})$$

$$\hat{a} = \arg\min_a L$$

- note

Logistic regression

$$\log\frac{p(1|\boldsymbol{x})}{p(-1|\boldsymbol{x})} = \boldsymbol{a}^T\boldsymbol{x}$$

$$p(1|\boldsymbol{x}) = \frac{e^{\boldsymbol{a}^T\boldsymbol{x}}}{1+e^{\boldsymbol{a}^T\boldsymbol{x}}}$$

$$p(-1|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{a}^T\boldsymbol{x}}}$$

$$L = -\sum_i \frac{1+y_i}{2}\boldsymbol{a}^T\boldsymbol{x}_i - \ln(1+e^{\boldsymbol{a}^T\boldsymbol{x}_i})$$

$$\hat{a} = \arg\min_a L$$

- note

$$p(1|\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{a}^T\boldsymbol{x}}}$$

$$p(-1|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{a}^T\boldsymbol{x}}}$$

$$
\begin{aligned}
p(y|\boldsymbol{x}) &= \frac{1}{1+e^{-y\boldsymbol{a}^T\boldsymbol{x}}} \qquad (y=\pm 1) \\
L(y_i,\gamma_i) &= -\ln[p(y_i|\boldsymbol{x}_i)] \\
&= \ln(1+e^{-y_i\gamma_i}) \qquad (\gamma_i=\boldsymbol{a}^T\boldsymbol{x}_i)
\end{aligned}
$$

正则化

$$L = -\sum_i \frac{1+y_i}{2}\boldsymbol{a}^T\boldsymbol{x}_i - \ln(1+e^{\boldsymbol{a}^T\boldsymbol{x}_i}) + \lambda\|\boldsymbol{a}\|_p$$

- $p=2$

$$\|\boldsymbol{a}\|_2 = \lambda\boldsymbol{a}^T\boldsymbol{a}$$

- p=1

$$\|\boldsymbol{a}\|_1 = \sum_i |a_i|$$

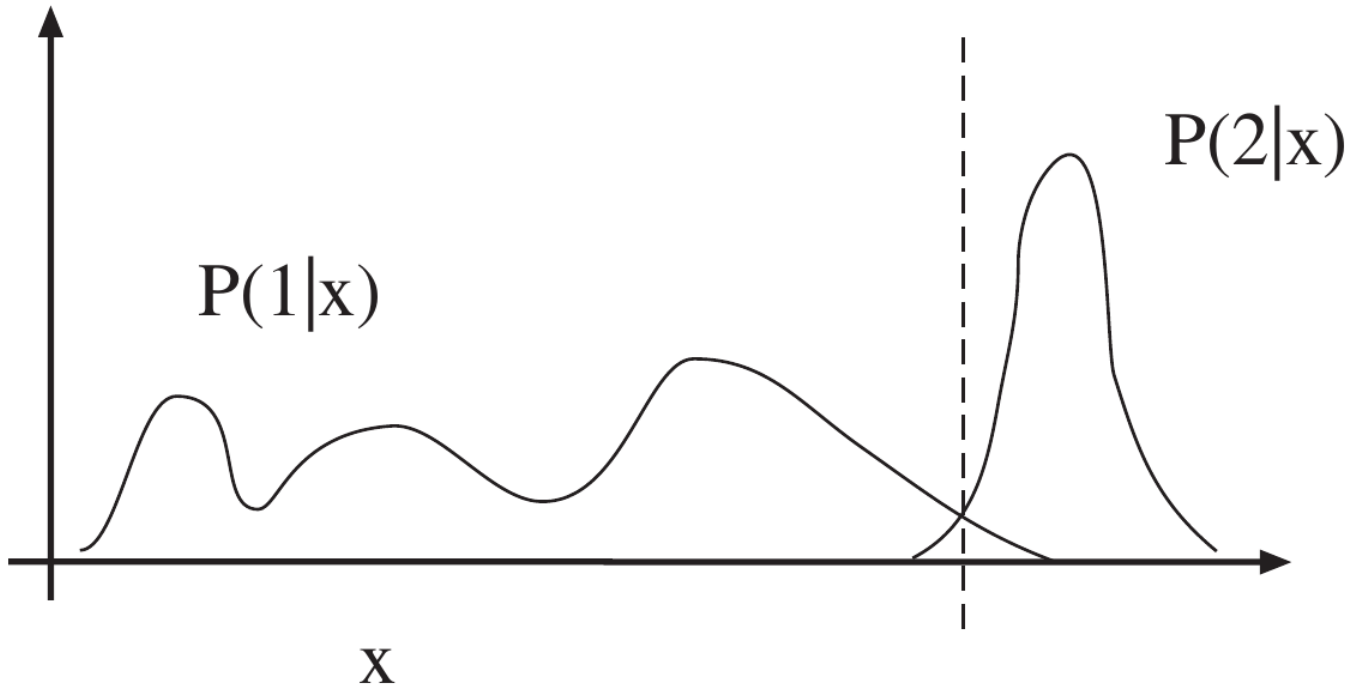


Figure 25.2. The figure shows posterior densities for two classes. The optimal decision boundary is shown as a dashed line. Notice that, while a normal density may provide rather a poor fit *to the posteriors*, the quality of the classifier it provides depends only on *how well it predicts the position of the boundaries.* In this case, assuming that the posteriors are normal may provide a fairly good classifier, because $P(2|x)$ looks normal, and the mean and covariance of $P(1|x)$ look as though they would predict the boundary in the right place.

正态类别条件分布

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \boldsymbol{x}_{k,i}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k - 1}\sum_{i=1}^{N_k} (\boldsymbol{x}_{k,i} - \boldsymbol{\mu}_k)(\boldsymbol{x}_{k,i} - \boldsymbol{\mu}_k)^T$$

选择

$$c = \arg\min_k \delta(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^2 - \Pr\{k\}$$

其中

$$\delta(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{2}((\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k))^{1/2}$$

k邻分类器

对于特征向量$\boldsymbol{x}$

- 确定距离$\boldsymbol{x}$最近的k个训练样例
- 确定k个样例中属于各类别的样例数
- 有最多样例的类别记作c,个数记作n
- $n > l$时$x \in c$，否则拒绝分类

估计、提高性能

- Cross Validation

  划分数据集，交替选择训练集与测试集

- Bootstrapping

  部分数据集用于训练，余下的测试，出错样例再次训练

皮肤像素分类

比较：

$$\frac{p(\boldsymbol{x}|\text{skin})p(\text{skin})}{p(\boldsymbol{x})}L(\text{skin} \to \text{not skin})$$

$$\frac{p(\boldsymbol{x}|\text{not skin})p(\text{not skin})}{p(\boldsymbol{x})}L(\text{not skin} \to \text{skin})$$

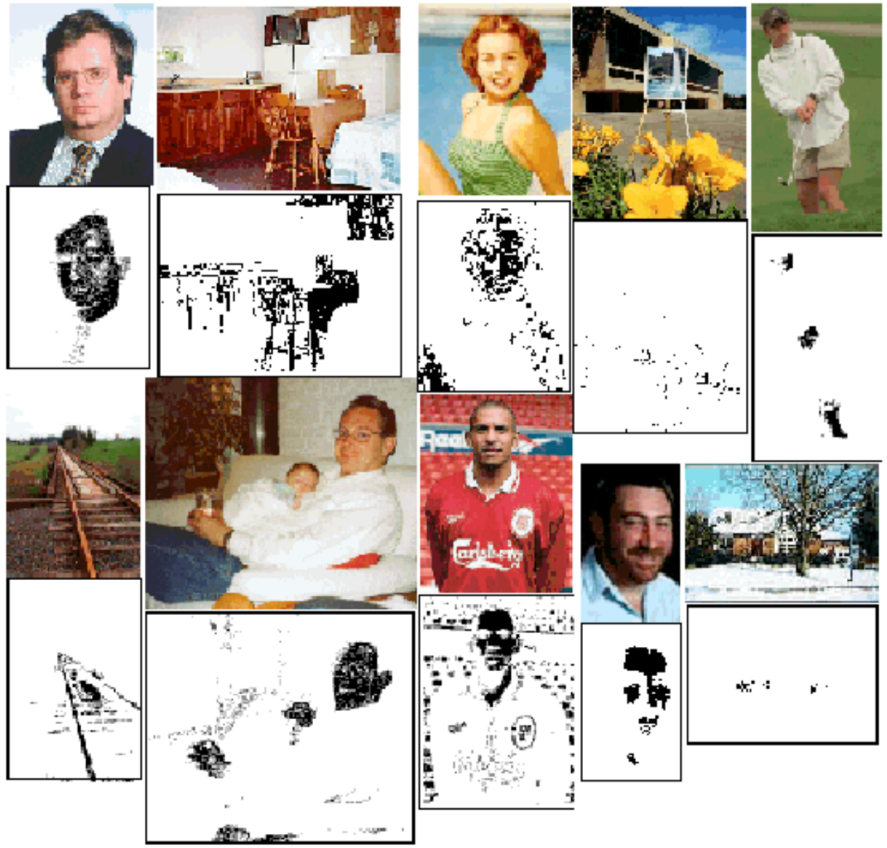


Figure 25.3. The figure shows a variety of images together with the output of the skin detector of Jones and Rehg applied to the image. Pixels marked black are skin pixels, and white are background. Notice that this process is relatively effective, and could certainly be used to focus attention on, say, faces and hands. *Figure from "Statistical color models with application to skin detection," M.J. Jones and J. Rehg, Proc. Computer Vision and Pattern Recognition, 1999* © *1999, IEEE*

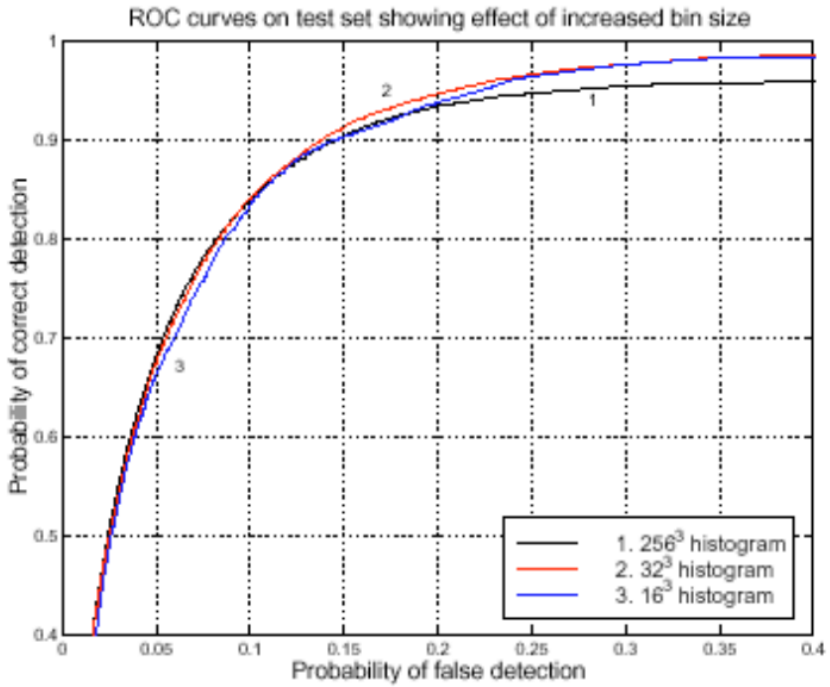


Figure 25.4. The receiver operating curve for the skin detector of Jones and Rehg. This plots the detection rate against the false negative rate for a variety of values of the parameter θ. A perfect classifier has an ROC that, on these axes, is a horizontal line at 100% detection. Notice that the ROC varies slightly with the number of boxes in the histogram. *Figure from "Statistical color models with application to skin detection," M.J. Jones and J. Rehg, Proc. Computer Vision and Pattern Recognition, 1999*

$$
\begin{aligned}
P(\text{image}|\text{face}) &= P(\text{label 1 at } (x1\,, y1), \ldots, \text{label } k \text{ at } (\text{xk}\,, \text{yk})|\text{face}) \\
&= P(\text{label 1 at } (x1\,, y1)|\text{face}) \cdots P(\text{label } k \text{ at } (\text{xk}\,, \text{yk}\,)|\text{face})
\end{aligned}
$$

$$
\begin{aligned}
v(\boldsymbol{x}_i) &= \boldsymbol{v}^T(\boldsymbol{x}_i - \boldsymbol{\mu}) \\
\text{var}(\boldsymbol{v}) &= \frac{1}{n-1} v(\boldsymbol{x}_i) v(\boldsymbol{x}_i)^T \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \boldsymbol{v}^T(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{v} \\
&= \boldsymbol{v}^T \Sigma \boldsymbol{v}
\end{aligned}
$$

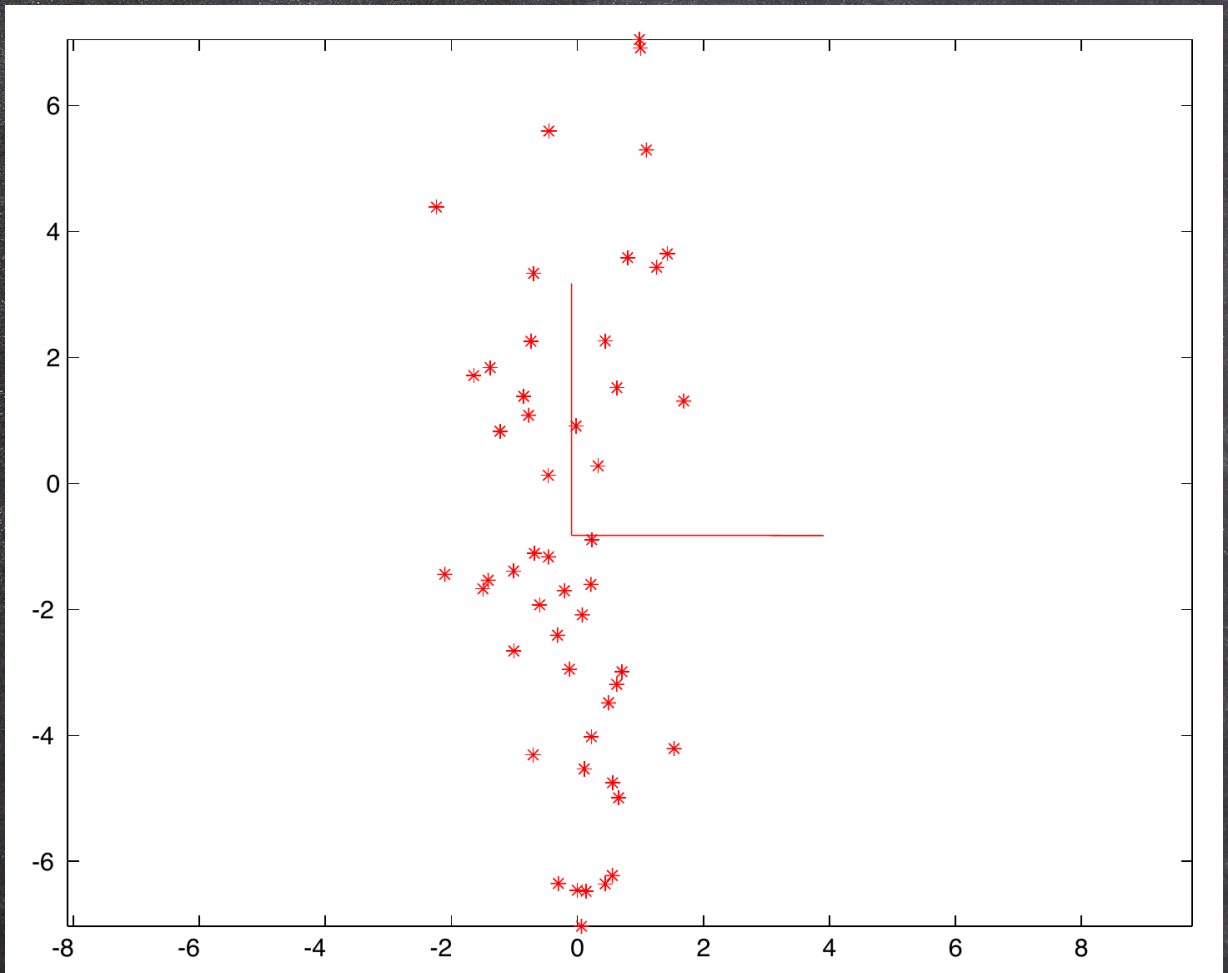


Figure 25.6. A data set which is well represented by a principal component analysis. The axes represent the directions obtained using PCA; the vertical axis is the first principal component, and is the direction in which the variance is highest.

Canonical Variates

$$\bar{\boldsymbol{\mu}} = \frac{1}{g}\sum_{j=1}^{g} \boldsymbol{\mu}_j$$

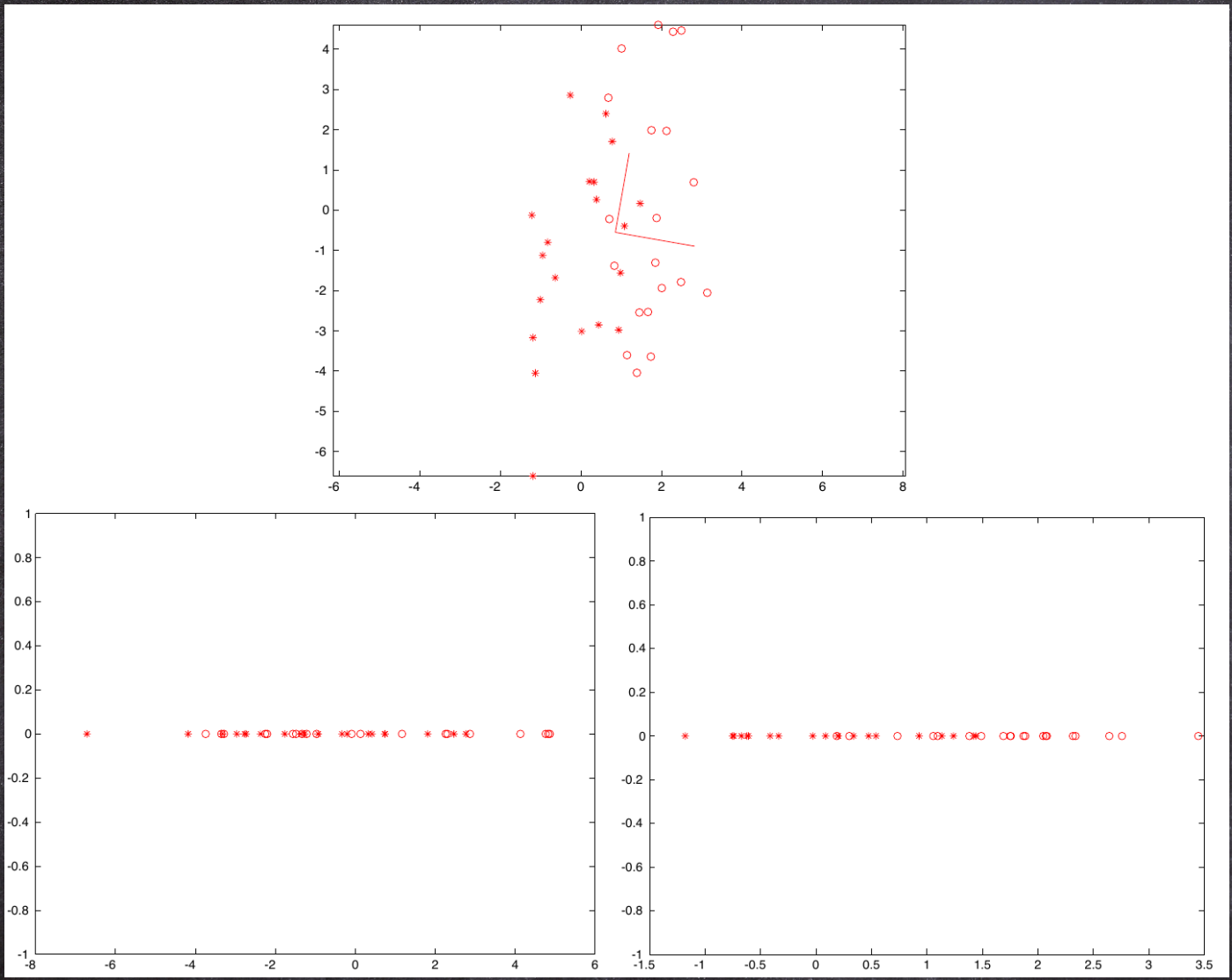
$$\mathcal{B} = \frac{1}{g-1}\sum_{j=1}^{g} (\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})^T$$

最大化

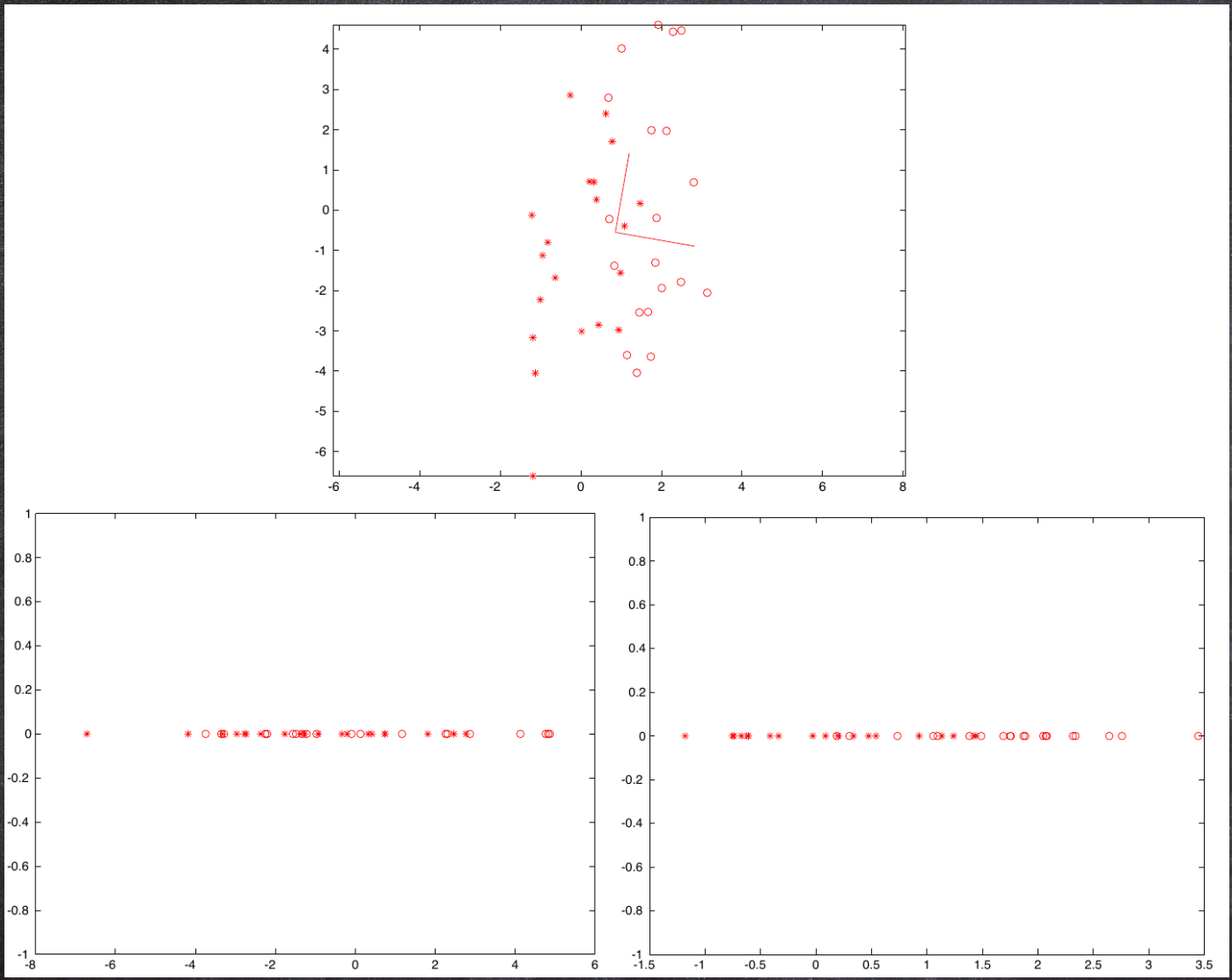
$$\frac{\boldsymbol{v}_1^T \mathcal{B} \boldsymbol{v}_1}{\boldsymbol{v}_1^T \Sigma \boldsymbol{v}_1}$$

得

$$\mathcal{B}\boldsymbol{v}_1 + \lambda \Sigma \boldsymbol{v}_1 = 0$$


4
3
2
1
0
-1
-2
-3
-4
-5
-6
-6
-4
-2
0
2
4
6
8
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-8
-6
-4
-2
0
2
4
6
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5

- Figure 25.10


4
3
2
1
0
-1
-2
-3
-4
-5
-6
-6
-4
-2
0
2
4
6
8
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-8
-6
-4
-2
0
2
4
6
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5

- Figure 25.10

Principal component analysis doesn't take into account the fact that there may be more than one class of item in a dataset. This can lead to significant problems. For a classifier, we would like to obtain a set of features that firstly reduces the number of features and secondly makes the difference between classes most obvious. For the data set on the top, one class is indicated by circles and the other by stars. PCA would suggest projection onto a vertical axis, which captures the variance in the dataset, but cannot be used to discriminate it, as we can see from the axes obtained by PCA, which are overlaid on the data set. The bottom row shows the projections onto those axes. On the bottom left, we show the projection onto the first principal component — which has higher variance, but separates the classes poorly — and on the bottom right, we show the projection onto the second principal component — which has significantly lower variance (look at the axes) and gives better separation.