第 5 章　特征选择与特征提取

在介绍非线性判别函数分类器和统计分类器识别方法之前，我们先来讨论一些与模式描述特征有关的问题。

通过前面的学习可以看出，无论是分类器设计还是聚类，依据的都是描述模式的识别特征。使用什么样的特征进行识别是与具体应用相关的，在绪论中我们用水果识别的例子说明了特征生成的过程，提取了颜色和形状两维特征；对于大多数实际问题来说，仅仅生成两维特征往往是不够的，识别需要的特征维数可能很高。例如，在手写数字识别中，如果每个字符的图像是大小为 28×28 的灰度图像，最直接的特征生成方法是将每个点的灰度值作为一维特征，这样就会得到 784 维的识别特征；在文本分类的应用中，一种常用的特征生成方法是以每个词语在文本中出现的频率作为识别特征，汉语、英语这些主要语言中常用词语的数量大多会超过 3 万个；在生物信息学的研究中，一项重要的工作是根据对生物基因组数据的分析寻找出导致某种疾病的遗传基因，像昆虫这样的低等生物一般会有数千个基因，而人类的基因数量则要超过 20000 个，如果以每个基因作为一维特征的话，我们就将面临着对一组高维特征矢量进行分类的问题。

识别特征维数的增高会给分类器的设计和学习带来很大的困难。一方面，高维特征增大了分类学习过程和识别过程计算和存储的复杂程度，降低了分类器的效率；另一方面，识别特征维数过大使得分类器过于复杂，这常常是一个更加严重的问题。

从本质上来讲本书所介绍的各种分类器学习过程都是利用统计学的方法，从训练样本中总结出能够区分不同类别的规律，提取出相关的信息。类别的区分信息大多是以分类器参数的形式进行描述的，例如在线性分类器中需要学习的参数是权值矢量。识别特征维数增加造成的分类器复杂主要体现在需要学习的分类器参数的增多，线性分类器权值矢量的维数随着特征维数 d 线性增长（特征维数加 1），后两章中将要学习的较为复杂的分类器，其参数随 d 增长的速度会更快。

分类器的学习过程实际上是一个利用训练样本估计参数的过程，根据统计学的知识我们知道，样本数量越多对统计量的估计就越准确；在样本数量一定的条件下，估计参数的数量越少准确度越高，而用少量样本估计过多的参数则是一个不可靠的过程。例如，已知某类样本来自于高斯分布，如果特征只有 1 维，那么使用 10 个训练样本所估计出的均值 μ 和方差 σ^2 具有一定的可信度；而当特征维数增大到 100 时，只用 10 个样本来估计 100 维的均值矢量 $\boldsymbol{\mu}$ 和 100×100 维的协方差矩阵 $\boldsymbol{\Sigma}$（共计 5150 个参数）则是完全无法接受的。实际分类问题的训练样本数量总是有限的，模式识别研究中将使用少量样本学习复杂分类器的问题形象地称为“维数的诅咒（Curse of Dimensionality）”。

本章我们将介绍一些在生成特征之后降低特征维数的方法，这些方法可以分为两大类：

特征选择和特征提取。

特征选择（Feature Selection）：所谓特征选择是指从原始生成的 d 维特征 $\mathbf{x}=(x_1,x_2,\cdots,x_d)^t$ 中挑选出 d' 个特征构成新的特征矢量 $\mathbf{x}'=(x_{i_1},x_{i_2},\cdots,x_{i_{d'}})^t$ 的过程，$d'<d$，$i_1,\cdots,i_{d'}\in\{1,\cdots,d\}$。特征选择的目的是要从原始的特征中挑选出对分类最有价值的一组特征，而抛弃掉与分类无关或对区分不同类别贡献很小的特征；

特征提取（Feature Extraction）：特征提取也是由原始的 d 维特征 $\mathbf{x}=(x_1,x_2,\cdots,x_d)^t$ 得到 d' 维特征的过程。但是与特征选择不同，这些特征不是从原始特征中直接挑选出来的，而是依据某种变换得到的。将经过提取降维之后的特征表示为 $\mathbf{y}=(y_1,y_2,\cdots,y_{d'})^t$，其中 $y_i=f_i(\mathbf{x})$，每一维新的特征都是由一个定义在原始特征 $\mathbf{x}$ 上的函数 f_i 映射得到的。

绪论中为了解决桃子和橘子的分类问题，我们分别提取了 RGB 三个分量的颜色特征 (x_1,x_2,x_3)，以及高度和宽度两个形状特征 (x_4,x_5)，这样就生成了一个原始的 5 维特征 $\mathbf{x}=(x_1,x_2,x_3,x_4,x_5)^t$。考虑到蓝色分量 x_3 在桃子和橘子图像中都比较少，对分类的作用很小，可以直接将其剔除，这实际上就是一个特征的选择过程；由于红色分量与绿色分量以及高度与宽度之间具有相关性，最终的 2 维识别特征 $\mathbf{y}=(y_1,y_2)^t$ 是分别由原始特征经过计算得到的，其中 $y_1=f_1(\mathbf{x})=x_2/x_1$，$y_2=f_2(\mathbf{x})=x_4/x_5$，这实际上是一个特征提取的过程。

在水果识别的例子中我们是通过人的观察完成的特征选择和提取，但对于复杂的分类问题来说就很难根据直观感觉实现特征的选择和提取了。例如，生物信息学对基因进行分析时，很难直观判断出哪些基因是与某种疾病直接相关的，在文本分类中也很难确定哪些词语能够区分不同类别的文本。在模式识别的研究中，我们希望能够依据一定的训练样本集，采用某些算法自动选择或提取出对分类有价值的特征从而降低特征的维数。

为了实现特征的选择和提取，首先需要有一个能够评价特征“价值”的准则，有了这样的准则才能够判断所选择或提取出来的特征是否对解决分类问题是有效的。对于特征选择来说，从 d 个原始特征中挑出 d' 个特征有很多种组合，计算每一种组合的有效性，然后找出其中对分类价值最大的一组往往是不可行的，这里需要研究的是如何能够利用有限的计算资源快速地找出一组“好的”特征；对于特征提取来说，需要研究的是如何找到一组“合理”的映射方式 $f_1(\mathbf{x}),\cdots,f_{d'}(\mathbf{x})$，能够将特征矢量由原始的 d 维空间映射到较低维数的 d' 维空间。

5.1 类别可分性判据

对于模式识别问题来说，一组特征的“价值”体现在使用这组特征构建的分类器是否能够很好地区分不同类别的样本，度量识别特征这种价值的指标一般称为**类别可分性判据**。

评价一组识别特征对类别是否具有可分性最直接的办法是使用这组特征设计和学习分类器，然后以分类的性能来衡量这组特征的优劣。然而很多分类器的学习算法都比较复杂，计算量较大，而且分类器的性能不仅决定于所使用的特征，也会受到很多其它因素的影响，例如不同的学习算法、分类器参数以及算法迭代初始值等。因此人们总是希望在学习分类器之前就能够依据训练样本集来度量特征对类别可分性的贡献，从而完成对特征的选择和提取。

在模式识别的研究过程中提出了很多以训练样本为基础度量特征可分性的方法，本节我们介绍两类简单和易于实现的类别可分性判据：基于距离的判据和基于散布矩阵的判据。

5.1.1 基于距离的可分性判据

从类别的可分性来看，自然是同一类别的样本相似性越大，不同类别的样本相似性越小对分类越有利，因此可以用样本之间的距离作为度量特征可分性的判据。

为了描述方便，将 c 个类别的样本集分别表示为 $D_1,\cdots,D_c$，其中 $D_i=\left\{\mathbf{x}_1^{(i)},\cdots,\mathbf{x}_{n_i}^{(i)}\right\}$，样本的上标表示所属类别，$n_i$ 为第 i 个类别的样本数；特征以集合的形式表示：$\mathcal{X}=\{x_1,\cdots x_d\}$。

类内距离

类内距离度量的是在特定特征集合 $\mathcal{X}$ 上同类别样本之间的相似程度。第 i 类样本集合中任意两个样本之间的均方距离：

$$d_i^2=\frac{1}{2n_i^2}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}d^2\left(\mathbf{x}_k^{(i)},\mathbf{x}_l^{(i)}\right) \tag{5.1}$$

所有类别样本总的均方距离：

$$J_{msd}\left(\mathcal{X}\right)=\sum_{i=1}^{c}P_i d_i^2 \tag{5.2}$$

其中 P_i 是第 i 个类别的先验概率，可以用第 i 类样本在全部样本中所占的比例来估计：

$$P_i\approx\frac{n_i}{n},\quad n=\sum_{j=1}^{c}n_j$$

当采用欧氏距离度量时，（5.2）式可以写成：

$$J_{msd}\left(\mathcal{X}\right)=\sum_{i=1}^{c}\frac{P_i}{2n_i^2}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{x}_l^{(i)}\right)^t\left(\mathbf{x}_k^{(i)}-\mathbf{x}_l^{(i)}\right) \tag{5.3}$$

可以证明，类内距离判据的简化计算方式为：

$$J_{msd}\left(\mathcal{X}\right)=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\boldsymbol{\mu}^{(i)}\right)^t\left(\mathbf{x}_k^{(i)}-\boldsymbol{\mu}^{(i)}\right) \tag{5.4}$$

其中 $\boldsymbol{\mu}^{(i)}$ 是第 i 类样本的均值：

$$\boldsymbol{\mu}^{(i)}=\frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{x}_k^{(i)}$$

类内距离判据 J_{msd} 度量的是在特征集合 $\mathcal{X}$ 上类内样本的聚集程度。

类间距离

类间距离度量的是不同类别样本之间的差异程度。第 i 个类别和第 j 个类别之间任意两个样本的均方距离：

$$d_{ij}^2=\frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}d^2\left(\mathbf{x}_k^{(i)},\mathbf{x}_l^{(j)}\right) \tag{5.5}$$

所有不同类别样本之间的均方距离：

$$J_{bsd}(\mathcal{X})=\frac{1}{2}\sum_{i=1}^{c}P_i\sum_{j=1,j\neq i}^{c}P_j\frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}d^2\left(\mathbf{x}_k^{(i)},\mathbf{x}_l^{(j)}\right) \tag{5.6}$$

当采用欧氏距离度量时，可以证明，总的类间均方距离有如下两种简化计算方式：

$$J_{bsd}(\mathcal{X})=\frac{1}{2}\sum_{i=1}^{c}P_i\sum_{j=1}^{c}P_j\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}^{(j)}\right)^t\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}^{(j)}\right) \tag{5.7a}$$

$$J_{bsd}(\mathcal{X})=\sum_{i=1}^{c}P_i\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}\right)^t\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}\right) \tag{5.7b}$$

其中 $\boldsymbol{\mu}$ 是所有类别样本的均值：

$$\boldsymbol{\mu}=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\mathbf{x}_k^{(i)}$$

5.1.2 基于散布矩阵的可分性判据

另一类可分性判据是定义在样本散布矩阵上的，类内散布矩阵 $\mathbf{S}_w$ 描述的是同类样本在特征空间中的分布情况：

$$\mathbf{S}_w=\sum_{i=1}^{c}P_i\mathbf{S}_i \tag{5.8}$$

其中：

$$\mathbf{S}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\boldsymbol{\mu}^{(i)}\right)\left(\mathbf{x}_k^{(i)}-\boldsymbol{\mu}^{(i)}\right)^t$$

$\mathbf{S}_i$ 是第 i 个类别的类内散布矩阵，$\mathbf{S}_w$ 是所有类别的类内散布矩阵。类间散布矩阵 $\mathbf{S}_b$ 描述的是不同类别样本在特征空间中的分布情况：

$$\mathbf{S}_b=\sum_{i=1}^{c}P_i\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}\right)\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}\right)^t \tag{5.9}$$

从（5.8）式和（5.9）式可以看出$\mathbf{S}_w$和$\mathbf{S}_b$均为$d\times d$的对称矩阵，对比（5.4）式和（5.7b）式还可以得到这样的结论：矩阵$\mathbf{S}_w$的主对角线元素之和为欧氏距离度量下的类内均方距离，而矩阵$\mathbf{S}_b$的主对角线元素之和是欧氏距离度量下的类间均方距离。矩阵$\mathbf{S}_w$和$\mathbf{S}_b$的非主对角线元素分别描述了同类样本和不同类样本对应特征对之间的相关程度。

除了类内和类间散布矩阵之外还可以定义所有样本的总体散布矩阵$\mathbf{S}_t$：

$$\mathbf{S}_t=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\boldsymbol{\mu}\right)\left(\mathbf{x}_k^{(i)}-\boldsymbol{\mu}\right)^t \tag{5.10}$$

实际上总体散布矩阵$\mathbf{S}_t$就是训练样本集$D=D_1\bigcup\cdots\bigcup D_c$的协方差矩阵，可以证明：

$$\mathbf{S}_t=\mathbf{S}_w+\mathbf{S}_b \tag{5.11}$$

由三个散布矩阵可以定义出很多可分性判据，常用的有：

$$J_1(\mathcal{X})=\mathrm{tr}\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right) \tag{5.12}$$

$$J_2(\mathcal{X})=\frac{\mathrm{tr}(\mathbf{S}_b)}{\mathrm{tr}(\mathbf{S}_w)} \tag{5.13}$$

$$J_3(\mathcal{X})=\frac{|\mathbf{S}_b|}{|\mathbf{S}_w|}=\left|\mathbf{S}_w^{-1}\mathbf{S}_b\right| \tag{5.14}$$

$$J_4(\mathcal{X})=\frac{|\mathbf{S}_t|}{|\mathbf{S}_w|} \tag{5.15}$$

其中$|\mathbf{A}|$表示矩阵$\mathbf{A}$的行列式值。

本节所介绍的类别可分性判据同第 3 章“聚类分析”中的聚类准则非常相似，两者都是评价样本集在一组特征上的区分程度。差别只是在于聚类分析中的样本集是无监督的，每个样本没有所属类别的信息，建立聚类准则的目的是要评价将样本集划分成不同的子集时，不同子集之间的区分程度；而在特征选择和提取中的样本集是有监督的，可分性判据评价的是这个样本集在不同特征子集上的区分程度。

【例 5.1】已知两类样本，计算 3 维特征中任意 2 维的类别可分性判据J_1。

$$\omega_1:\mathbf{x}_1=(0,0,0)^t,\mathbf{x}_2=(1,0,0)^t,\mathbf{x}_3=(2,2,1)^t,\mathbf{x}_4=(1,1,0)^t$$

$$\omega_2:\mathbf{x}_5=(0,0,1)^t,\mathbf{x}_6=(0,2,0)^t,\mathbf{x}_7=(0,2,1)^t,\mathbf{x}_8=(1,1,1)^t$$

解：首先计算第 1、2 维特征上每个类别的均值和样本的总体均值：

$$\boldsymbol{\mu}_1=\frac{1}{4}\sum_{i=1}^{4}\mathbf{x}_i=(1.00,0.75)^t,\boldsymbol{\mu}_2=\frac{1}{4}\sum_{i=5}^{8}\mathbf{x}_i=(0.25,1.25)^t$$

$$\boldsymbol{\mu}=\frac{1}{8}\sum_{i=1}^{8}\mathbf{x}_i=(0.625,1.000)^t$$

计算类内散布矩阵：

$$\mathbf{S}_w=\frac{1}{2}\left[\frac{1}{4}\sum_{i=1}^{4}(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^t+\frac{1}{4}\sum_{i=5}^{8}(\mathbf{x}_i-\boldsymbol{\mu}_2)(\mathbf{x}_i-\boldsymbol{\mu}_2)^t\right]=\begin{pmatrix}0.3438 & 0.2188\\ 0.2188 & 0.6875\end{pmatrix}$$

计算类间散布矩阵：

$$\mathbf{S}_b=\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu})(\boldsymbol{\mu}_1-\boldsymbol{\mu})^t+\frac{1}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu})(\boldsymbol{\mu}_2-\boldsymbol{\mu})^t=\begin{pmatrix}0.1406 & -0.0938\\ -0.0938 & 0.0625\end{pmatrix}$$

因此：

$$\mathbf{S}_w^{-1}\mathbf{S}_b=\begin{pmatrix}0.6218 & -0.4145\\ -0.3342 & 0.2228\end{pmatrix}$$

第 1、2 维特征的类别可分性判据：

$$J_1(x_1,x_2)=\mathrm{tr}\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right)=0.8446$$

同样的过程可以计算出第 1、3 维特征和第 2、3 维特征的可分性判据：

$$J_1(x_1,x_3)=1.9268\ ,\quad J_1(x_2,x_3)=0.3750$$

显然如果要从 3 个原始特征中选择 2 维，第 1、3 维特征是最佳的选择。从图 5.1 也可以看出，两类样本在 x_1-x_3 平面上的投影混叠最小。■

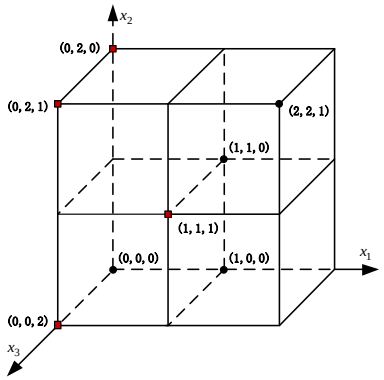

图 5.1 例 5.1 的样本分布

5.2 特征选择

特征选择的目的是要从原始的特征集合 $\mathcal{X}$ 中挑选出一组最有利于分类的特征 $\mathcal{X}'$，由于类别可分性判据可以评价挑选出的一组特征对于分类问题的有效性，因此特征选择实际上就是一个对某种选定的可分性判据 J 的优化：

$$\mathcal{X}'=\arg\max_{\tilde{\mathcal{X}}\subset\mathcal{X}} J\left(\tilde{\mathcal{X}}\right) \tag{5.16}$$

其中原始特征集合 $\mathcal{X}$ 中包含 d 个特征，$\mathcal{X}'$ 中包含 $d'<d$ 个特征，$\tilde{\mathcal{X}}$ 是任意包含 d' 个元素的 $\mathcal{X}$ 的子集。

求解（5.16）式优化问题的一个简单思路是用可分性判据 J 分别评价每一个特征，然后根据判据值的大小对特征重新排序，使得：

$$J(x_1) \geq J(x_2) \geq \cdots \geq J(x_d) \tag{5.17}$$

选择判据值最大的前 d' 个特征作为特征选择的结果：$\mathcal{X}' = \{x_1, x_2, \cdots, x_{d'}\}$。然而，遗憾的是这种方法选择出的特征集合 $\mathcal{X}'$ 并不能保证是（5.16）式的最优解，因为在这个过程中并没有考虑各个特征之间的相关性。只有当特征之间相互独立时才能够保证解的最优性，而当特征之间存在着相关性时，判据值最大的 d' 个特征组合在一起不能保证可分性是最优的。

另一个求解（5.16）式优化问题的简单思路是对所有 $\tilde{\mathcal{X}} \subset \mathcal{X}$ 的特征组合进行穷举，计算每一种组合的判据值，选择出最优组合。穷举法可以保证选出最优的特征组合，然而需要以巨大的计算复杂度为代价，因此只具有理论上的可行性。从 d 个特征中选择 d' 个特征共有 $C_d^{d'}$ 种组合，例 5.1 中由 3 个特征选择出 2 个特征只需要考虑 3 种组合情况，而当需要从 100 个特征中选择 10 个时，组合数则变为 $C_{100}^{10} = 17310309456440$。

5.2.1 分支定界法

分支定界法是一种能够减小穷举法计算复杂度的最优特征组合搜索算法，但是它依赖于类别可分性判据的一个重要性质--单调性，即对于两个特征子集 $\mathcal{X}_1$ 和 $\mathcal{X}_2$ 来说：

$$\mathcal{X}_1 \subset \mathcal{X}_2 \Rightarrow J(\mathcal{X}_1) \leq J(\mathcal{X}_2) \tag{5.18}$$

也就是说如果从某个特征集合中去除一个特征将会减小判据值，如果增加一个特征则会增大判据值。单调性并不是所有的类别可分性判据都具有的性质，可以验证，类内、类间均方距离 J_{msd} 、 J_{bsd} 以及 J_1 和 J_3 满足单调性，而 J_2 和 J_4 则不满足。只有当可分性判据满足单调性时，分支定界法才能够保证搜索到最优的特征组合。

下面先通过一个例子来说明分支定界法的过程。假设我们要从原始的 $d = 6$ 维特征 $\mathcal{X} = \{x_1, x_2, \cdots, x_6\}$ 中选择出 $d' = 2$ 的特征组合，使得某个满足单调性的可分性判据 J 最大。

分支定界法首先需要自顶向下构建一个如图 5.2 所示的搜索树，树的每个节点是一个特征组合，根节点对应全部的 6 维原始特征；每个节点的子节点对应的是从父节点特征集中删除一个特征的子集，节点旁的数字表示被删除的特征，例如根节点的左子节点 C 代表的就是删除特征 x_1 后的子集；每向下一级删除一个特征，由 6 维特征选择 2 维特征共需删除 4 个特征，因此树的深度是 4，每个叶节点对应着所有可能的 $C_6^2 = 15$ 种特征组合。

图 5.2 的搜索树是非对称的，这样可以保证生成的叶节点不会出现重复的特征组合。这一点可以由如下事实看出，从 6 个特征中选择出 2 个，需要删除 4 个特征，将这 4 个特征的下标编号记为(f_1，f_2，f_3，f_4)。实际上我们真正关心的是删除的是哪 4 个特征，而不关心删除的先后次序，因此不妨规定 $f_1<f_2<f_3<f_4$，这样就可以列出对应的 15 种特征删除方式：(1234)，(1235)，(1236)，(1245)，(1246)，(1256)，(1345)，(1346)，(1356)，(1456)，(2345)，

(2346)，(2356)，(2456)，(3456)。观察一下图 5.2 中由根节点到每一个叶节点路径上节点旁的数字就会发现，每一条路径刚好就是这 15 种特征删除方式中的一种。例如根节点到最右边的叶节点 A 的路径对应的是(3456)，相应的叶节点则是被选择出来的 2 维特征组合 (x_1, x_2)。

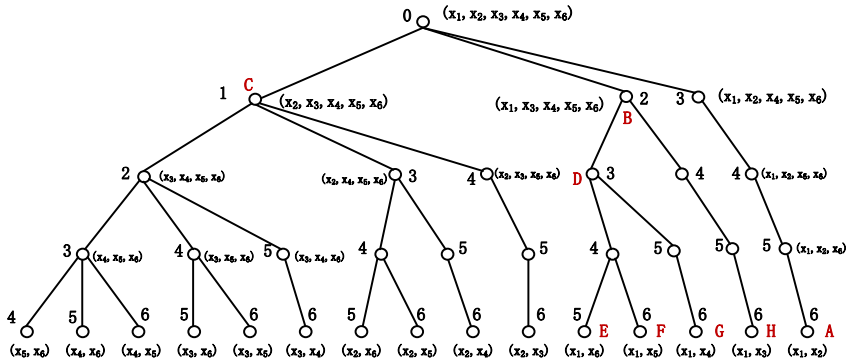

图 5.2 分支定界搜索树

有了搜索树，下面来看如何实现快速搜索最优的特征组合。分支定界法按照由右至左深度优先的方式进行搜索，首先计算节点 A 对应特征组合的判据值 $J(A)$ 作为当前的最优结果；回溯搜索至节点 B，计算判据值 $J(B)$；如果 $J(B) < J(A)$，由于 E，F，G，H 均为 B 的后继节点，对应的特征组合均为 B 的子集，根据判据 J 的单调性，有：

$$J(E), J(F), J(G), J(H) \le J(B) < J(A)$$

因此在 B 的后继节点中不可能存在最优节点，无需继续搜索；回溯搜索节点 C，如果 $J(C) < J(A)$，那么节点 C 的后继节点中也不会存在最优节点，由此可以判断节点 A 为最优节点，最优的特征组合为 $\{x_1, x_2\}$。在这个过程中我们只是计算了 A，B，C 三个节点的判据值，明显要少于计算全部 15 个叶节点的判据值。

当然，这是一种理想的情况，如果 $J(B) > J(A)$，那么需要继续向下搜索节点 H 和节点 D。下面给出分支定界法的完整过程：

分支定界法

- 初始化：根据原始特征维数 d 和选择特征维数 d' 构建搜索树，设置界值 $B = 0$；
- 从右向左分支定界搜索：
  - 如果当前节点没有分支，则向下搜索，直到叶节点为止。计算叶节点代表的特征集合的可分性判据，如果大于界值 B，则将 B 替换为这个判据值，并记录这个特征集合，作为当前的最优选择；向上回溯，直到有节点存在未搜索过的分支为止，按照从右向左的顺序搜索其子节点；
  - 如果当前节点有分支，则计算当前节点代表特征集合的可分性判据，小于界值 B，则中止该节点向下的搜索；否则按照从右向左的顺序搜索其子节点
- 输出：最优特征集合

分支定界算法存在两个主要问题，首先，算法是否能够搜索到最优的特征组合依赖于所采用的类别可分性判据是否具有单调性，不具有单调性的可分性判据分支定界算法不能保证得到最优的特征选择结果；其次，分支定界算法的计算复杂度是不确定的，与最优解分支所在位置有关，如果最优解分支在最右端并且根节点的子节点判据值均小于最优解，则搜索效率最高；如果每个分支的可分性判据都大于其左端分支的可分性判据，那么需要计算搜索树上所有节点的判据值，实际的计算复杂度会超过穷举法。在图 5.2 的例子中有可能需要计算 24 次可分性判据，多于穷举法的 15 次计算。

5.2.2 次优搜索算法

由于分支定界法对可分性判据有着严格的单调性要求，而且当原始的特征维数 d 很大时，搜索到最优解需要的计算量仍然是可观的，往往无法满足需要。因此更多的实际问题不再追求寻找到最优的特征组合，转而采用某种次优搜索算法，选择出一组比较好的特征。

最简单的次优搜索算法是使用可分性判据单独评价每一个特征的优劣，按照（5.17）式排序选择出 d' 个单独最优的特征，这种方法计算简单，只需要计算 d 次可分性判据。正如前面分析的一样，这种方法完全忽略了特征之间的相关性，得到的结果往往不能令人满意。

顺序前进法（Sequential Forward Selection, SFS）

顺序前进法也称作自下而上的搜索方法。从一个空集开始每次向选择的特征集合中加入一个特征，直到特征集合中包含 d' 个特征为止，每次选择加入特征的原则是将其加入特征集后能够使得可分性判据最大。顺序前进法的过程可以用如下算法表示：

顺序前进法

- 初始化：原始特征集合 $\mathcal{X}$，设置选择特征集合 $\mathcal{X}'=\Phi$；
- 循环直到 $\mathcal{X}'$ 中包含 d' 个特征为止：
  - 计算将任意未被选择的特征加入 $\mathcal{X}'$ 后的可分性判据值：$J\left(\mathcal{X}'\cup\{x_i\}\right)$，$\forall x_i\in\mathcal{X}-\mathcal{X}'$；
  - 寻找最优特征：$x'=\arg\max_{x_i\in\mathcal{X}-\mathcal{X}'} J\left(\mathcal{X}'\cup\{x_i\}\right)$；
  - 将最优特征加入选择特征集合：$\mathcal{X}'=\mathcal{X}'\cup\{x'\}$
- 输出：特征集合 $\mathcal{X}'$

顺序前进法每一轮迭代只需计算将每一个未被选择的特征加入 $\mathcal{X}'$ 之后的判据值，因此选择出 d' 个特征需要计算判据值的次数为：

$$\sum_{i=0}^{d'-1}(d-i)=\frac{d'(2d-d'+1)}{2} \tag{5.19}$$

顺序后退法 (Sequential Backward Selection, SBS)

顺序后退法也称作自上而下的搜索方法。同顺序前进法的过程相反，首先开始于整个

特征集$\mathcal{X}$，每一轮从特征集中选择一个最差的特征删除，选择特征的原则是将其删除之后使得特征集合的判据值下降的最小。顺序后退法的过程为：

顺序后退法

- 初始化：原始特征集合$\mathcal{X}$，设置选择特征集合$\mathcal{X}'=\mathcal{X}$；
- 循环直到$\mathcal{X}'$中包含d'个特征为止：
  - 计算将任意一个$\mathcal{X}'$中元素删除之后的可分性判据值：$J\left(\mathcal{X}'-\{x_i\}\right)$，$\forall x_i\in\mathcal{X}'$；
  - 寻找最优的删除特征：$x'=\arg\max_{x_i\in\mathcal{X}'} J\left(\mathcal{X}'-\{x_i\}\right)$；
  - 将选择的特征移出集合：$\mathcal{X}'=\mathcal{X}'-\{x'\}$
- 输出：特征集合$\mathcal{X}'$

顺序后退法每一轮迭代需要计算将$\mathcal{X}'$中的每个元素删除之后的判据值，直到$\mathcal{X}'$中剩余d'个元素为止，需要迭代$d-d'$次，因此判据值的计算次数为：

$$\sum_{i=0}^{d-d'-1}(d-i)=\frac{(d-d')(d+d'+1)}{2} \tag{5.20}$$

广义顺序前进(后退)法（Generalized Sequential Forward(Backward) Selection, GSFS, GSBS）

顺序前进和顺序后退法每次向特征集合中增加或删除 1 个特征，而广义的顺序前进和后退法则是每次增加或删除r个特征。

广义顺序前进法每一轮迭代需要从未被选择的特征集合$\mathcal{X}-\mathcal{X}'$中寻找最优的r个特征的组合加入$\mathcal{X}'$，而广义顺序后退法则是每一轮迭代需要从$\mathcal{X}'$的元素中寻找删除r个特征的最优组合，如果共进行了k轮迭代的话，判据值的计算次数为：

$$\sum_{i=0}^{k-1}C_{d-i\times r}^{r}=\frac{1}{r!}\times\sum_{i=0}^{k-1}\frac{(d-i\times r)!}{(d-i\times r-r)!} \tag{5.21}$$

一般来说广义顺序前进、后退法的计算量都要大于顺序前进和顺序后退法。但是由于每次选择特征时都是寻找r个特征的最优组合，因此在一定程度上考虑了特征之间的统计相关性，所以优化的结果一般要好于每次选择一个特征的顺序前进或后退法。

增 l – 减 r 法（$l-r$法）

在顺序前进法中，一旦某个特征被加入到选择的特征集合$\mathcal{X}'$，就不会被删除了；而在顺序后退法中，某个特征被从$\mathcal{X}'$中删除，则不会再被加入了。这实际上对搜索最优的特征组合是不利的，因为在选择这些特征时只考虑了它与当前在$\mathcal{X}'$中的特征之间的相关性，以及增加或删除之后的判据值大小，而没有考虑之后加入或删除$\mathcal{X}'$中某些特征时的情况。增 l – 减 r 法允许对特征选择的过程进行回溯，先采用顺序前进法向选择特征集合$\mathcal{X}'$加入l个特征，然后采用顺序后退法从$\mathcal{X}'$中删除r个特征（$l>r$），循环这个过程直到$\mathcal{X}'$中包含d'个特征为止。

增 l － 减 r 法

- 初始化：设置选择特征集合 $\mathcal{X}'=\Phi$；
- 循环直到 $\mathcal{X}'$ 中包含 d' 个特征为止：
  - 调用顺序前进法 l 次，向 $\mathcal{X}'$ 中添加 l 个特征；
  - 调用顺序后退法 r 次，从 $\mathcal{X}'$ 中删除 r 个特征；
- 输出：特征集合 $\mathcal{X}'$

回溯的过程也可以按照相反的顺序进行，从全部的特征集合开始 $\mathcal{X}'=\mathcal{X}$，先采用顺序后退法从 $\mathcal{X}'$ 中删除 r 个特征，然后采用顺序前进法将 l（$l<r$）个特征加入 $\mathcal{X}'$，直到 $\mathcal{X}'$ 中包含 d' 个特征为止。

5.3 特征提取

特征提取和特征选择的目的都是要降低特征的维数，不同的是特征选择是在原始特征 $\mathbf{x}=(x_1,\cdots,x_d)^t$ 中挑选出 d' 个特征使得某种类别可分性判据最优，而特征提取则是要构造一组定义在矢量 $\mathbf{x}$ 上的函数 $f_1(\mathbf{x}),\cdots,f_{d'}(\mathbf{x})$。一般来说 $f_i(\mathbf{x})$ 可以是任意形式的函数，在这里我们讨论其中一类最简单的形式--线性函数，介绍两种最常用的线性特征提取方法。

5.3.1 主成分分析（Principle Component Analysis, PCA）

对于一个样本集合来说，原始特征描述每个样本使用了 d 个特征，降低维数之后只使用了 d' 个特征。一般来说随着数据量（特征数量）的减少，样本中的信息会有所丢失，新的特征对样本的描述也会存在一定的误差，主成分分析方法就是从尽量减少信息损失的角度来实现特征降维的。

主成分分析算法和推导

我们知道，样本集合中的每一个样本都对应着特征空间中的一个点，同样的一个样本点在不同坐标系下对应着不同的矢量，例如图 5.3 中某个样本对应着特征空间中的点 A，在以 O 为坐标原点 $\{\mathbf{v}_1,\mathbf{v}_2\}$ 为基矢量的原坐标系下 A 点对应的矢量是 $\mathbf{x}$，而在 O’为原点 $\{\mathbf{e}_1,\mathbf{e}_2\}$ 为基矢量的新坐标系下对应的矢量是 $\mathbf{x}'$，如果新坐标系的原点 O'在原坐标系下对应的矢量是 $\boldsymbol{\mu}$，显然有如下关系：

$$\mathbf{x}=\boldsymbol{\mu}+\mathbf{x}' \tag{5.22}$$

这个关系虽然是从 2 维空间中得到的，在高维空间中同样成立。

分别将矢量 $\mathbf{x}$ 和 $\mathbf{x}'$ 写成两个坐标系下的坐标形式：$\mathbf{x}=(x_1,\cdots,x_d)^t$，$\mathbf{x}'=(a_1,\cdots,a_d)^t$，参考附录 A.3，（5.22）式可以表示为：

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{i=1}^{d} a_i \mathbf{e}_i \tag{5.23}$$

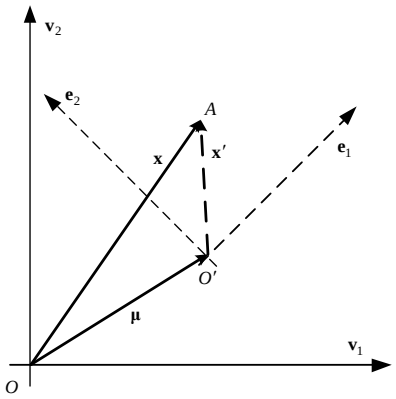

图 5.3 同一个样本在不同坐标系下的表示

在新坐标系下矢量，$\mathbf{x}'$ 的元素可以由原坐标系下的矢量 $\mathbf{x}$ 以及 $\boldsymbol{\mu}$ 和基矢量 $\{\mathbf{e}_1,\cdots,\mathbf{e}_d\}$ 计算得到：

$$a_i = \mathbf{e}_i^t(\mathbf{x}-\boldsymbol{\mu}),\quad i=1,\cdots,d \tag{5.24}$$

同样也可以根据（5.23）式由新坐标下的矢量 $\mathbf{x}'$ 来恢复原矢量 $\mathbf{x}$，不会存在任何误差。然而如果我们只保留新坐标系下 $d'<d$ 个元素，然后用保留的 d' 个元素来恢复原坐标系下的 d 维矢量：

$$\hat{\mathbf{x}} = \boldsymbol{\mu} + \sum_{i=1}^{d'} a_i \mathbf{e}_i \tag{5.25}$$

显然 $\hat{\mathbf{x}}$ 只是对 $\mathbf{x}$ 的近似，用 $\hat{\mathbf{x}}$ 来代替 $\mathbf{x}$ 就会出现一定的误差，误差的大小同新坐标系的位置、基矢量的方向以及保留哪些特征有关。

在主成分分析方法中，新坐标系的原点选择在训练样本集 $D=\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ 的均值矢量 $\boldsymbol{\mu}$ 上，然后寻找一组最优的基矢量 $\{\mathbf{e}_1,\cdots,\mathbf{e}_d\}$，使得在只保留前 d' 个元素的条件下，由新的坐标根据（5.25）式恢复样本集 D 的均方误差最小，即求解如下的优化问题：

$$\min_{\mathbf{e}_1,\cdots,\mathbf{e}_d} J(\mathbf{e}_1,\cdots,\mathbf{e}_d) = \frac{1}{n}\sum_{k=1}^{n}\left\|\mathbf{x}_k - \hat{\mathbf{x}}_k\right\|^2 \tag{5.26}$$

其中 $\hat{\mathbf{x}}_k$ 是根据（5.24）式将 $\mathbf{x}_k$ 由原坐标系变换到新坐标系下，然后再根据（5.25）式只使用前 d' 个特征恢复的近似矢量。如果用 a_{ki} 表示第 k 个样本在新坐标系下的第 i 维特征，由（5.23）式和（5.25）式可以得到：

$$\mathbf{x}_k - \hat{\mathbf{x}}_k = \sum_{i=d'+1}^{d} a_{ki}\mathbf{e}_i \tag{5.27}$$

代入到（5.26）式：

$$
\begin{aligned}
J\left(\mathbf{e}_1,\cdots,\mathbf{e}_d\right) &= \frac{1}{n}\sum_{k=1}^{n}\left\|\mathbf{x}_k-\hat{\mathbf{x}}_k\right\|^2 = \frac{1}{n}\sum_{k=1}^{n}\left\|\sum_{i=d'+1}^{d}a_{ki}\mathbf{e}_i\right\|^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=d'+1}^{d}a_{ki}\mathbf{e}_i\right)^t\left(\sum_{i=d'+1}^{d}a_{ki}\mathbf{e}_i\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=d'+1}^{d}a_{ki}^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=d'+1}^{d}\left[\mathbf{e}_i^t\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\right]\left[\mathbf{e}_i^t\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\right]^t \\
&= \sum_{i=d'+1}^{d}\mathbf{e}_i^t\left[\frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^t\right]\mathbf{e}_i
\end{aligned}
$$

其中第 2 行到第 3 行利用了$\{\mathbf{e}_1,\cdots,\mathbf{e}_d\}$是新坐标系的基矢量，因此构成了一个标准正交系：

$$
\mathbf{e}_i^t\mathbf{e}_j=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases},\quad i,j=1,\cdots,d
$$

而第 3 行到第 4 行则是基于如下事实：a_{ki}是一个标量，它的转置与其自身相等，并且有（5.24）式成立，因此$a_{ki}=\mathbf{e}_i^t\left(\mathbf{x}_k-\boldsymbol{\mu}\right)=\left[\mathbf{e}_i^t\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\right]^t$。如果定义矩阵：

$$
\boldsymbol{\Sigma}=\frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^t
$$

恰好是样本集D的协方差矩阵，则（5.26）式的优化问题变为：

$$
\min_{\mathbf{e}_1,\cdots,\mathbf{e}_d} J\left(\mathbf{e}_1,\cdots,\mathbf{e}_d\right)=\sum_{i=d'+1}^{d}\mathbf{e}_i^t\boldsymbol{\Sigma}\mathbf{e}_i \tag{5.28}
$$

仔细观察（5.28）式会发现，直接求解这个优化问题是没有意义的。由于$\boldsymbol{\Sigma}$是半正定矩阵，因此当$\mathbf{e}_1,\cdots,\mathbf{e}_d=\mathbf{0}$时取得最小值 0，显然零矢量并不能作为基矢量，导致这样结果的原因在于优化（5.28）式时没有约束$\mathbf{e}_1,\cdots,\mathbf{e}_d$的长度。主成分分析在求解新坐标系的基矢量时优化的是如下的约束问题：

$$
\min_{\mathbf{e}_1,\cdots,\mathbf{e}_d} J\left(\mathbf{e}_1,\cdots,\mathbf{e}_d\right)=\sum_{i=d'+1}^{d}\mathbf{e}_i^t\boldsymbol{\Sigma}\mathbf{e}_i \tag{5.29}
$$

约束：

$$
\left\|\mathbf{e}_i\right\|^2=1,\quad i=1,\cdots,d
$$

有约束优化问题可以通过构造 Lagrange 函数转化为无约束问题（参见附录 B.6）：

$$
L\left(\mathbf{e}_1,\cdots,\mathbf{e}_d,\lambda_1,\cdots,\lambda_d\right)=\sum_{i=d'+1}^{d}\mathbf{e}_i^t\boldsymbol{\Sigma}\mathbf{e}_i-\sum_{i=d'+1}^{d}\lambda_i\left(\mathbf{e}_i^t\mathbf{e}_i-1\right) \tag{5.30}
$$

对每一个基矢量 $\mathbf{e}_j$ 求导数：

$$\frac{\partial L(\mathbf{e}_1,\cdots,\mathbf{e}_d,\lambda_1,\cdots,\lambda_d)}{\partial \mathbf{e}_j}=2\boldsymbol{\Sigma}\mathbf{e}_j-2\lambda_j\mathbf{e}_j=0 \tag{5.31}$$

其中利用到了 $\boldsymbol{\Sigma}$ 为对称矩阵的事实。由此得到优化问题（5.29）的解应满足：

$$\boldsymbol{\Sigma}\mathbf{e}_j=\lambda_j\mathbf{e}_j \tag{5.32}$$

显然，使得上式成立的 λ_j 和 $\mathbf{e}_j$ 分别为矩阵 $\boldsymbol{\Sigma}$ 的特征值和对应的特征矢量。由此我们可以得到这样的结论：如果我们希望将一个样本集合 D 中的 d 维特征矢量在一个新的坐标系下只用 d' 个特征进行表示，那么应该将新坐标系的坐标原点放在 D 的均值 $\boldsymbol{\mu}$ 的位置，而以集合 D 的协方差矩阵的特征矢量 $\mathbf{e}_1,\cdots,\mathbf{e}_d$ 作为基矢量，这样可以保证只用保留的 d' 维特征恢复原矢量时均方误差最小。

通过这样的方式可以得到一个最优的新坐标系，注意到 $\boldsymbol{\Sigma}$ 是一个 $d\times d$ 的矩阵，存在 d 个特征值和特征矢量，现在的问题是我们只希望保留新坐标系中的 d' 个坐标，应该保留哪些坐标才能够保证恢复出的 d 维特征矢量的均方误差最小？回到优化问题（5.29），将（5.32）式代入优化函数：

$$J(\mathbf{e}_1,\cdots,\mathbf{e}_d)=\sum_{i=d'+1}^{d}\mathbf{e}_i^t\boldsymbol{\Sigma}\mathbf{e}_i=\sum_{i=d'+1}^{d}\lambda_i\mathbf{e}_i^t\mathbf{e}_i=\sum_{i=d'+1}^{d}\lambda_i \tag{5.33}$$

要使得 $J(\mathbf{e}_1,\cdots,\mathbf{e}_d)$ 最小，只需要选择 $\lambda_{d'+1},\cdots,\lambda_d$ 是 $\boldsymbol{\Sigma}$ 最小的 $d-d'$ 个特征值。这里需要注意一点，在整个推导过程中我们约定的是要保留新坐标系下前 d' 个特征，而放弃掉后面的 $d-d'$ 个特征，因此在新的坐标系下应该选择保留的是 $\boldsymbol{\Sigma}$ 最大的 d' 个特征值对应的特征矢量作为新坐标系的基矢量。

主成分分析算法可以用如下的过程描述：

主成分分析

- 输入样本集合 D，计算均值矢量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$；
- 计算矩阵 $\boldsymbol{\Sigma}$ 的特征值和特征矢量，按照特征值由大到小排序；
- 选择前 d' 个特征矢量作为列矢量构成矩阵 $\mathbf{E}=(\mathbf{e}_1\quad\mathbf{e}_2\quad\cdots\quad\mathbf{e}_{d'})$；
- d 维特征矢量 $\mathbf{x}$ 可以转换为 d' 维矢量 $\mathbf{x}'$：$\mathbf{x}'=\mathbf{E}^t(\mathbf{x}-\boldsymbol{\mu})$。

在模式识别中，主成分分析方法通常被用于降低特征的维数，采用上述过程就可以将所有的训练样本，以及需要识别的样本由 d 维特征矢量转换为 d' 维特征矢量；在某些应用中可能还希望由降维之后的矢量 $\mathbf{x}'$ 来恢复原矢量 $\mathbf{x}$，通过下面公式的计算可以达到这个目的：

$$\hat{\mathbf{x}}=\mathbf{E}\mathbf{x}'+\boldsymbol{\mu} \tag{5.34}$$

【例 5.2】样本集 D 中包含 8 个样本，采用主成分分析的方法将 2 维特征降为 1 维：

$$\mathbf{x}_1=(10,1)^t,\mathbf{x}_2=(9,0)^t,\mathbf{x}_3=(10,-1)^t,\mathbf{x}_4=(11,0)^t$$
$$\mathbf{x}_5=(0,9)^t,\mathbf{x}_6=(1,10)^t,\mathbf{x}_7=(0,11)^t,\mathbf{x}_8=(-1,10)^t$$

解：计算样本的均值矢量

$$\boldsymbol{\mu}=\frac{1}{8}\sum_{i=1}^{8}\mathbf{x}_i=\frac{1}{8}\left[\begin{pmatrix}10\\1\end{pmatrix}+\begin{pmatrix}9\\0\end{pmatrix}+\begin{pmatrix}10\\-1\end{pmatrix}+\begin{pmatrix}11\\0\end{pmatrix}+\begin{pmatrix}0\\9\end{pmatrix}+\begin{pmatrix}1\\10\end{pmatrix}+\begin{pmatrix}0\\11\end{pmatrix}+\begin{pmatrix}-1\\10\end{pmatrix}\right]=\begin{pmatrix}5\\5\end{pmatrix}$$

计算协方差矩阵

$$\begin{aligned}\boldsymbol{\Sigma}&=\frac{1}{8}\sum_{i=1}^{8}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^t\\&=\frac{1}{8}\left\{\left[\begin{pmatrix}5\\-4\end{pmatrix}\begin{pmatrix}5&-4\end{pmatrix}+\begin{pmatrix}4\\-5\end{pmatrix}\begin{pmatrix}4&-5\end{pmatrix}+\begin{pmatrix}5\\-6\end{pmatrix}\begin{pmatrix}5&-6\end{pmatrix}+\begin{pmatrix}6\\-5\end{pmatrix}\begin{pmatrix}6&-5\end{pmatrix}\right]+\right.\\&\qquad\left.\begin{pmatrix}-5\\4\end{pmatrix}\begin{pmatrix}-5&4\end{pmatrix}+\begin{pmatrix}-4\\5\end{pmatrix}\begin{pmatrix}-4&5\end{pmatrix}+\begin{pmatrix}-5\\6\end{pmatrix}\begin{pmatrix}-5&6\end{pmatrix}+\begin{pmatrix}-6\\5\end{pmatrix}\begin{pmatrix}-6&5\end{pmatrix}\right\}\\&=\begin{pmatrix}25.5&-25\\-25&25.5\end{pmatrix}\end{aligned}$$

求协方差矩阵 $\boldsymbol{\Sigma}$ 特征值和特征矢量，（5.32）式可以写成关于特征矢量 $\mathbf{e}$ 的方程组形式：

$$(\boldsymbol{\Sigma}-\lambda\mathbf{I})\mathbf{e}=\mathbf{0}\qquad(5.35)$$

这是一个齐次线性方程组，其中 $\mathbf{I}$ 是 $d\times d$ 维的单位矩阵。方程组有解的条件是系数矩阵 $\boldsymbol{\Sigma}-\lambda\mathbf{I}$ 的行列式值等于 0：

$$|\boldsymbol{\Sigma}-\lambda\mathbf{I}|=\begin{vmatrix}25.5-\lambda&-25\\-25&25.5-\lambda\end{vmatrix}=(25.5-\lambda)^2-25^2=0$$

这样可以解得两个特征值：

$$\lambda_1=50.5,\quad\lambda_2=0.5$$

将 λ_1 代入（5.35）式方程组：

$$\begin{pmatrix}-25&-25\\-25&-25\end{pmatrix}\begin{pmatrix}e_1\\e_2\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

因此有 $e_1=-e_2$。不妨设 $e_1=1$，就得到了对应 λ_1 的特征矢量：$\mathbf{e}_1=(1,-1)^t$。$\mathbf{e}_1$ 不是单位矢量，作为基矢量需要标准化：

$$\mathbf{e}_1=\frac{\mathbf{e}_1}{\|\mathbf{e}_1\|}=\begin{pmatrix}\sqrt{2}/2\\-\sqrt{2}/2\end{pmatrix}$$

同样可以得到对应于 λ_2 的特征矢量：

$$\mathbf{e}_2=\begin{pmatrix}\sqrt{2}/2\\\sqrt{2}/2\end{pmatrix}$$

由于 $\lambda_1>\lambda_2$，因此应该选择 $\mathbf{e}_1$ 作为主分量，D 中样本在新坐标系下降维之后的结果为：

$$x_1' = \mathbf{e}_1^t(\mathbf{x}_1 - \boldsymbol{\mu}) = \left(\sqrt{2}/2, -\sqrt{2}/2\right)\begin{pmatrix}5\\-4\end{pmatrix} = 9\sqrt{2}/2$$

$$x_2' = 9\sqrt{2}/2, \quad x_3' = 11\sqrt{2}/2, \quad x_4' = 11\sqrt{2}/2$$

$$x_5' = -9\sqrt{2}/2, \quad x_6' = -9\sqrt{2}/2, \quad x_7' = -11\sqrt{2}/2, \quad x_8' = -11\sqrt{2}/2$$

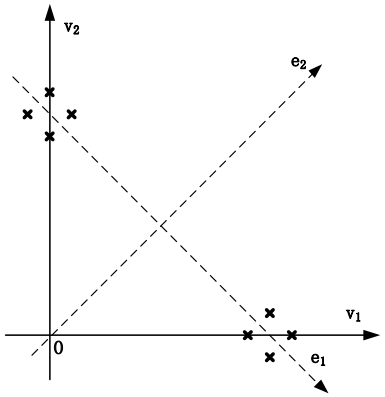

图 5.4 例 2.1 示意图

■

主成分分析的相关问题

主成分分析的过程非常简单，只要计算训练样本集协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值和特征矢量，就可以得到一组基矢量从而构成新的坐标系，然后将特征矢量向这些基矢量上投影即可以达到降维的目的。在这个过程中我们可能会产生这样一些疑问，对于任意一个方阵来说都存在特征值和特征矢量，但是特征值和特征矢量的元素有可能是复数，那么复数的特征值如何按照大小排序？复矢量如何成为新的坐标系的基矢量？在求解基矢量的优化问题中只约束解矢量是单位矢量，并没有对它们之间是否正交进行约束，那么主成分分析得到的新坐标系是否还是直角坐标系？

对于样本集的协方差矩阵 $\boldsymbol{\Sigma}$ 这样一个实对称矩阵来说，存在一些特殊的性质（参见附录 A.2），可以证明实对称矩阵的特征值都是实数，并且由于 $\boldsymbol{\Sigma}$ 是半正定矩阵，因此这些特征值都是大于等于 0 的实数，特征矢量的每个元素也都是实数，不会出现复矢量的情况。

同样，$d \times d$ 维的实对称矩阵 $\boldsymbol{\Sigma}$ 存在 d 个特征矢量，而且这些特征矢量之间是相互正交的（相同特征值对应的特征矢量可以进行正交化处理），因此主成分分析得到的新坐标系是一个直角坐标系。

我们还可以从另一个方面来理解主成分分析的结果：这样的特征变换消除了样本集 D 在各个特征之间的相关性，新的坐标系下不同特征之间是不相关的。由于协方差矩阵的主对角线元素是特征的方差，而非主对角线元素就是不同特征之间的相关系数，因此我们只需证明经过特征变换之后样本集 D 的协方差矩阵是对角阵即可证明主成分分析消除了特征之间的相关性。

首先计算新坐标系下样本的均值：

$$\boldsymbol{\mu}' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}^t\left(\mathbf{x}_i - \boldsymbol{\mu}\right) = \mathbf{E}^t\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i - \boldsymbol{\mu}\right) = \mathbf{0} \tag{5.36}$$

得到这样的结果并不奇怪，因为主成分分析是将新坐标系的原点移到了 $\boldsymbol{\mu}$ 的位置。再计算新的坐标系下样本集 D 的协方差矩阵：

$$\begin{aligned}
\boldsymbol{\Sigma}' &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i' - \boldsymbol{\mu}'\right)\left(\mathbf{x}_i' - \boldsymbol{\mu}'\right)^t \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{E}^t\left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right]\left[\mathbf{E}^t\left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right]^t \\
&= \mathbf{E}^t\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^t\right]\mathbf{E} \\
&= \mathbf{E}^t\boldsymbol{\Sigma}\mathbf{E} \\
&= \begin{pmatrix} \mathbf{e}_1^t \\ \vdots \\ \mathbf{e}_{d'}^t \end{pmatrix}\boldsymbol{\Sigma}\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_{d'} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{e}_1^t\boldsymbol{\Sigma}\mathbf{e}_1 & \mathbf{e}_1^t\boldsymbol{\Sigma}\mathbf{e}_2 & \cdots & \mathbf{e}_1^t\boldsymbol{\Sigma}\mathbf{e}_{d'} \\ \mathbf{e}_2^t\boldsymbol{\Sigma}\mathbf{e}_1 & \mathbf{e}_2^t\boldsymbol{\Sigma}\mathbf{e}_2 & \cdots & \mathbf{e}_2^t\boldsymbol{\Sigma}\mathbf{e}_{d'} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}_{d'}^t\boldsymbol{\Sigma}\mathbf{e}_1 & \mathbf{e}_{d'}^t\boldsymbol{\Sigma}\mathbf{e}_2 & \cdots & \mathbf{e}_{d'}^t\boldsymbol{\Sigma}\mathbf{e}_{d'} \end{pmatrix}
\end{aligned}$$

其中最后一步使用了分块矩阵的乘法。由于基矢量 $\mathbf{e}_i$ 和 $\mathbf{e}_j$ 是单位正交的，因此：

$$\mathbf{e}_i^t\boldsymbol{\Sigma}\mathbf{e}_j = \lambda_j\mathbf{e}_i^t\mathbf{e}_j = \begin{cases} \lambda_i, & i = j \\ 0, & i \neq j \end{cases} \tag{5.37}$$

这样就得到：

$$\boldsymbol{\Sigma}' = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{d'} \end{pmatrix} \tag{5.38}$$

主成分分析的结果还可以从几何的角度来理解，以 2 维特征为例，如果样本集 D 中的 n 个样本大致分布在一个椭圆形的区域内，例如图 5.5 的情形。主成分分析将新坐标系的坐标原点建立在椭圆的中心 $\boldsymbol{\mu}$ 处，最大特征值对应的基矢量 $\mathbf{e}_1$ 是在椭圆的长轴方向（一般也称为主分量），而另一个基矢量 $\mathbf{e}_2$ 则是椭圆的短轴方向。如果在新的坐标下选择 1 维特征表示样本显然应该选择 $\mathbf{e}_1$ 轴，因为在这个轴上的样本分布更接近于原始特征。d 维空间中的情形也是一样的，如果样本分布在一个 d 维的椭球之内，那么新坐标系的基矢量刚好对应着 d 个主轴的方向，对应特征值的大小表示的是相应椭球主轴的长短。

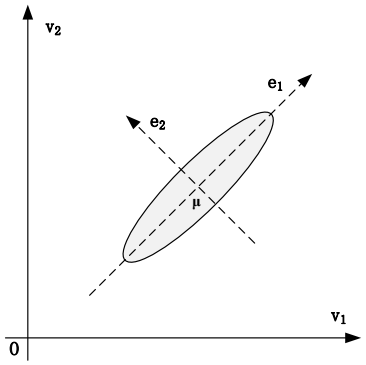

图 5.5 主成分分析的几何解释

使用主成分分析的方法降维还需要考虑的一个问题是对于训练样本集 D 来说，在新的坐标系下应该保留多少特征合适，即如何选择参数 d'？回答这个问题需要回到（5.33）式来看，主成分分析只用 d' 个新特征来表示原始的 d 维特征，这样必然会带来误差。均方误差的大小是被舍弃的 $d-d'$ 维新特征对应的特征值之和，如果被舍弃的特征值均为 0，那么就不会引起误差，舍弃的特征值越小则带来的误差越小。常用的一种 d' 参数的选择方法是将所有特征值按照由大到小排序之后计算累加值，以累加值与所有特征值的总和之比超过 95%为原则选择 d'，这种做法可以理解为我们希望降维之后能够保留样本集 D 的信息超过 95%，或者降维所带来的误差不超过 5%：

$$d' = \arg\min_{1\le k\le d}\left[\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{d}\lambda_i} \ge 95\%\right] \qquad (5.39)$$

下面 2 段 Matlab 代码分别实现的是主成分分析中的主分量提取和特征降维过程：

函数名称：PCA
参数：X--样本矩阵（n×d 矩阵），ratio--特征值累加和占总和的比例
返回值：E--基矢量矩阵（d×d’矩阵，每列一个矢量），mu--均值矢量（行矢量）
函数功能：主成分分析提取主分量

```
function [E, mu] = PCA( X, ratio )

d = size(X,2);                                  %计算均值和协方差矩阵
mu = mean(X);
Sigma = cov(X);

[V,L] = eigs( Sigma, d );                       %求特征值和特征矢量
Lamda = diag(L);
AccLamda = cumsum(Lamda);                       %计算累加特征值
t = AccLamda(d) * ratio;
dd = find( (AccLamda(2:d)>=t) & (AccLamda(1:d-1)<t) ) + 1;          %计算保留特征数 d'
E = V(:,1:dd);
```

函数名称：PCADR

参数：X--样本矩阵（n×d 矩阵），E--基矢量矩阵（d×d’矩阵，每列一个矢量），mu--均值矢量（行矢量）

返回值：Y--降维之后的样本矩阵（n×d’矩阵）

函数功能：主成分分析降维

```
function Y = PCADR(X, E, mu )

n = size(X,1);
Y = (X-repmat(mu,n,1))*E;
```

5.3.2 基于 Fisher 准则的可分性分析（Fisher Discriminant Analysis, FDA）

主成分分析的目标是要消除特征之间的相关性，而没有考虑样本集中样本的类别属性，因此是一种无监督学习方法。而模式识别的目标是要利用降维之后的样本设计和学习分类器，更关心的是降维之后是否能够保留不同类别样本之间的可分性信息，两者的目的是有差异的。我们先来回顾一下例 5.2，如果训练样本集 D 中的 8 个样本分别属于两个类别：

$$\omega_1:\left\{\mathbf{x}_1=(10,1)^t,\mathbf{x}_4=(11,0)^t,\mathbf{x}_6=(1,10)^t,\mathbf{x}_7=(0,11)^t\right\}$$

$$\omega_2:\left\{\mathbf{x}_2=(9,0)^t,\mathbf{x}_3=(10,-1)^t,\mathbf{x}_5=(0,9)^t,\mathbf{x}_8=(-1,10)^t\right\}$$

使用主成分分析的方法降为 1 维特征，那么所有样本都需要向 $\mathbf{e}_1$ 轴投影。从图 5.6 可以看出，两个类别的 8 个样本在原 2 维空间中是线性可分的，而当投影到 $\mathbf{e}_1$ 轴之后则是完全不可分的，如果在主成分分析方法中选择对应特征值较小的 $\mathbf{e}_2$ 轴进行投影反倒能够保留类别的可分性信息。这个例子说明在主成分分析方法中认为不重要的特征维度可能恰恰包含着对于分类来说重要的信息。

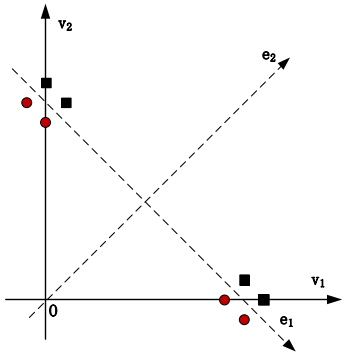

图 5.6 主成分分析方法中类别可分性信息的丢失

基于 Fisher 准则的可分性分析也是一种线性特征降维方法，有时被称为线性可分性分析（Linear Discriminant Analysis, LDA），它的出发点是要使得经过降维之后的特征能够尽量多地保留类别之间的可分性信息，换句话说就是要使得经过降维之后的样本集合具有最大的类别可分性。

基于 Fisher 准则的可分性分析算法与推导

下面先从一种简单的情况入手来研究这个问题，我们将两个类别的样本向一条通过坐标原点的直线上投影，也就是用 1 维特征来表示 d 维矢量，希望在 1 维空间中两类样本的可分性最大。从图 5.7 可以看出在不同方向的直线上，两类样本的可分性是不同的，如果想要找到一个最优的投影直线方向，首先需要对 1 维空间中样本的可分性进行度量。

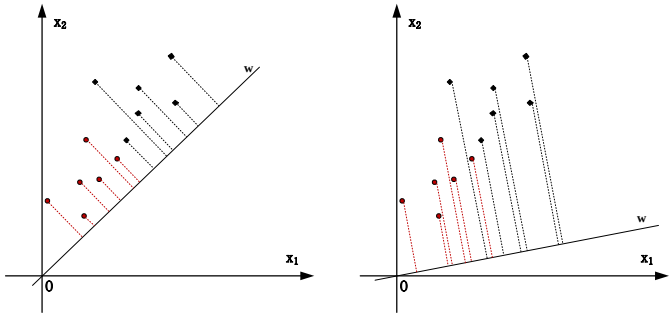

图 5.7 二维模式在一维空间的投影

假设两类问题的样本集为：$D_1=\left\{\mathbf{x}_1^{(1)},\cdots,\mathbf{x}_{n_1}^{(1)}\right\}$，$D_2=\left\{\mathbf{x}_1^{(2)},\cdots,\mathbf{x}_{n_2}^{(2)}\right\}$，投影直线的单位矢量为 $\mathbf{w}$，d 维空间的矢量 $\mathbf{x}$ 在这条直线上的投影为一个标量：

$$y=\mathbf{w}^t\mathbf{x} \tag{5.40}$$

两类样本集经过投影之后成为标量集：$D_1\to\mathcal{Y}_1=\left\{y_1^{(1)},\cdots,y_{n_1}^{(1)}\right\}$，$D_2\to\mathcal{Y}_2=\left\{y_1^{(2)},\cdots,y_{n_2}^{(2)}\right\}$。参照本章 5.1 节的可分性判据，不同类样本的分散程度越大，同类样本的聚集程度越高则类别之间的可分性越强。在 1 维空间中可以用两个类别样本均值之差的平方 $\left(\tilde{\mu}_1-\tilde{\mu}_2\right)^2$ 来度量两类样本的分散程度；而样本类内的离散程度可以用样本的方差之和来度量 $\tilde{s}_1^2+\tilde{s}_2^2$。综合考虑类内的聚集程度和类间的分散程度，可以建立如下的 Fisher 准则：

$$J\left(\mathbf{w}\right)=\frac{\left(\tilde{\mu}_1-\tilde{\mu}_2\right)^2}{\tilde{s}_1^2+\tilde{s}_2^2} \tag{5.41}$$

Fisher 准则函数的值越大，类别的可分性则越强。下面写出准则函数 $J\left(\mathbf{w}\right)$ 关于投影直线方向矢量 $\mathbf{w}$ 的显式表达式，首先计算投影之后类别的均值 $\tilde{\mu}$：

$$\tilde{\mu}_i=\frac{1}{n_i}\sum_{y\in\mathcal{Y}_i}y=\frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{w}^t\mathbf{x}=\mathbf{w}^t\boldsymbol{\mu}_i,\quad i=1,2 \tag{5.42}$$

投影之后两类均值之差的平方可以表示为：

$$\left(\tilde{\mu}_1-\tilde{\mu}_2\right)^2=\left(\mathbf{w}^t\boldsymbol{\mu}_1-\mathbf{w}^t\boldsymbol{\mu}_2\right)^2=\mathbf{w}^t\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^t\mathbf{w}=\mathbf{w}^t\mathbf{S}_b\mathbf{w} \tag{5.43}$$

其中 $\mathbf{S}_b=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t$ 是 5.1 节公式（5.9）所定义的类间散布矩阵（假设两类的先验概率相等）。类似的：

$$\begin{aligned}\tilde{s}_i^2&=\sum_{y\in\mathcal{Y}_i}(y-\tilde{\mu}_i)^2\\&=\sum_{\mathbf{x}\in D_i}\left(\mathbf{w}^t\mathbf{x}-\mathbf{w}^t\boldsymbol{\mu}_i\right)^2\\&=\sum_{\mathbf{x}\in D_i}\mathbf{w}^t(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^t\mathbf{w}\\&=\mathbf{w}^t\mathbf{S}_i\mathbf{w}\end{aligned}$$

其中 $\mathbf{S}_i=\sum_{\mathbf{x}\in D_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^t$ 类似于公式（5.8）所定义的类内散布矩阵（只相差一个比例系数）。总的方差：

$$\tilde{s}_1^2+\tilde{s}_2^2=\mathbf{w}^t\mathbf{S}_1\mathbf{w}+\mathbf{w}^t\mathbf{S}_2\mathbf{w}=\mathbf{w}^t(\mathbf{S}_1+\mathbf{S}_2)\mathbf{w}=\mathbf{w}^t\mathbf{S}_w\mathbf{w}\tag{5.44}$$

（5.43）、（5.44）式代入 Fisher 准则，可以得到如下优化问题：

$$\max_{\mathbf{w}}J(\mathbf{w})=\frac{(\tilde{\mu}_1-\tilde{\mu}_2)^2}{\tilde{s}_1^2+\tilde{s}_2^2}=\frac{\mathbf{w}^t\mathbf{S}_b\mathbf{w}}{\mathbf{w}^t\mathbf{S}_w\mathbf{w}}\tag{5.45}$$

上式也被称为是 Rayleigh 商的优化问题。实际上这个问题存在着无穷多个解，因为如果 $\mathbf{w}^*$ 是一个最优解的话，对于任意的 $a\neq 0$，$a\mathbf{w}^*$ 同样是最优解。我们真正关心的是投影矢量 $\mathbf{w}$ 的方向，而不关心它的长度（可以规格化为单位矢量），因此可以通过适当调整 $\|\mathbf{w}\|$ 使得 Fisher 准则的分母 $\mathbf{w}^t\mathbf{S}_w\mathbf{w}$ 等于一个常数 C。这样我们就得到了一个有约束的优化问题：

$$\max_{\mathbf{w}}J(\mathbf{w})=\mathbf{w}^t\mathbf{S}_b\mathbf{w}\tag{5.46}$$

约束：

$$\mathbf{w}^t\mathbf{S}_w\mathbf{w}=C$$

构造 Lagrange 函数转化为无约束优化：

$$L(\mathbf{w})=\mathbf{w}^t\mathbf{S}_b\mathbf{w}-\lambda\left(\mathbf{w}^t\mathbf{S}_w\mathbf{w}-C\right)\tag{5.47}$$

对 $\mathbf{w}$ 求导：

$$\frac{\partial L(\mathbf{w})}{\partial\mathbf{w}}=2\mathbf{S}_b\mathbf{w}-2\lambda\mathbf{S}_w\mathbf{w}=0$$

因此有：

$$\mathbf{S}_b\mathbf{w}=\lambda\mathbf{S}_w\mathbf{w}\tag{5.48}$$

满足上式的 λ 和 $\mathbf{w}$ 称为关于 $\mathbf{S}_b$ 和 $\mathbf{S}_w$ 的广义特征值和特征矢量，如果 $\mathbf{S}_w$ 的逆矩阵存在的话，有：

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{w}=\lambda\mathbf{w}\tag{5.49}$$

λ 和 $\mathbf{w}$ 分别是矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值和对应的特征矢量。与主成分分析一样，$\mathbf{S}_w^{-1}\mathbf{S}_b$ 是一

个 $d\times d$ 的方阵，存在 d 个特征值和 d 个特征矢量，哪一个特征矢量使得 Fisher 准则取得最大值？将满足（5.48）式的一个特征矢量 $\mathbf{w}_i$ 代入（5.45）式：

$$J(\mathbf{w})=\frac{\mathbf{w}_i^t\mathbf{S}_b\mathbf{w}_i}{\mathbf{w}_i^t\mathbf{S}_w\mathbf{w}_i}=\frac{\lambda_i\mathbf{w}^t\mathbf{S}_w\mathbf{w}_i}{\mathbf{w}_i^t\mathbf{S}_w\mathbf{w}_i}=\lambda_i \tag{5.50}$$

由此可见最大特征值对应的特征矢量是使得 Fisher 准则取得最大值的方向矢量。通过这样一个推导过程我们可以得出如下结论：两个类别的样本向一条直线上投影，当直线的方向矢量 $\mathbf{w}$ 为矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 最大特征值对应的特征矢量时，可以使得投影之后样本在 1 维空间中具有最大的可分性。

下面将这个问题推广到 c 个类别的样本向 d' 个方向上的投影，即将 d 维特征降维为 d' 个特征，使得降维之后的样本具有最大的可分性。首先定义矩阵：

$$\mathbf{S}_w=\sum_{i=1}^{c}\mathbf{S}_i\ ,\quad \mathbf{S}_b=\sum_{i=1}^{c}n_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^t \tag{5.51}$$

其中 $\boldsymbol{\mu}$ 为所有样本的均值矢量。可以证明使得 Fisher 准则最大的 d' 个投影矢量是对应于矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 最大 d' 个特征值的特征矢量。

基于 Fisher 准则的可分性分析可以用如下过程描述：

基于 Fisher 准则的可分性分析

- 根据（5.51）式计算矩阵 $\mathbf{S}_b$ 和 $\mathbf{S}_w$；
- 计算矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值和特征矢量，特征值由大到小排序；
- 选择前 d' 个特征矢量作为列矢量构成矩阵 $\mathbf{E}=(\mathbf{e}_1\quad\mathbf{e}_2\quad\cdots\quad\mathbf{e}_{d'})$；
- d 维特征矢量 $\mathbf{x}$ 可以转换为 d' 维矢量 $\mathbf{x}'$：$\mathbf{x}'=\mathbf{E}^t\mathbf{x}$。

【例 5.3】现有 3 个类别的样本，采用基于 Fisher 准则的可分性分析方法将 2 维特征降为 1 维。

$$\omega_1:\mathbf{x}_1=(1,3)^t,\mathbf{x}_2=(1,4)^t,\mathbf{x}_3=(3,0)^t,\mathbf{x}_4=(3,1)^t$$
$$\omega_2:\mathbf{x}_5=(3,6)^t,\mathbf{x}_6=(3,7)^t,\mathbf{x}_7=(5,5)^t,\mathbf{x}_8=(5,4)^t$$
$$\omega_3:\mathbf{x}_9=(8,5)^t,\mathbf{x}_{10}=(9,9)^t,\mathbf{x}_{11}=(9,5)^t,\mathbf{x}_{12}=(10,9)^t$$

解：首先计算各类样本的均值和总体均值：

$$\boldsymbol{\mu}=\frac{1}{12}\sum_{i=1}^{12}\mathbf{x}_i=(5,4.83)^t\ ,\quad \boldsymbol{\mu}_1=\frac{1}{4}\sum_{i=1}^{4}\mathbf{x}_i=(2,2)^t$$
$$\boldsymbol{\mu}_2=\frac{1}{4}\sum_{i=5}^{8}\mathbf{x}_i=(4,4.5)^t\ ,\quad \boldsymbol{\mu}_3=\frac{1}{4}\sum_{i=9}^{12}\mathbf{x}_i=(9,7)^t$$

计算矩阵 $\mathbf{S}_b$ 和 $\mathbf{S}_w$：

$$\mathbf{S}_b=\sum_{i=1}^{3}4(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^t=\begin{pmatrix}104 & 66\\ 66 & 52.7\end{pmatrix}$$

$$\mathbf{S}_w=\sum_{i=1}^{4}(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^t+\sum_{i=5}^{8}(\mathbf{x}_i-\boldsymbol{\mu}_2)(\mathbf{x}_i-\boldsymbol{\mu}_2)^t+\sum_{i=9}^{12}(\mathbf{x}_i-\boldsymbol{\mu}_3)(\mathbf{x}_i-\boldsymbol{\mu}_3)^t=\begin{pmatrix}10 & -6\\ -6 & 31\end{pmatrix}$$

因此：

$$\mathbf{S}_w^{-1}\mathbf{S}_b=\begin{pmatrix}13.2 & 8.62\\ 4.69 & 3.37\end{pmatrix}$$

计算矩阵$\mathbf{S}_w^{-1}\mathbf{S}_b$的特征值和特征矢量：

$$\lambda_1=16.33，\ \mathbf{e}_1=(0.94,0.33)^t$$

$$\lambda_2=0.25，\ \mathbf{e}_2=(-0.55,0.83)^t$$

表 5.1 例 5.3 样本在 FDA 两个方向上的投影值

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
$\mathbf{e}_1$	1.96	2.30	2.82	3.16	4.86	5.20	6.40	6.06	9.22	11.5	10.2	12.5
$\mathbf{e}_2$	1.94	2.78	-1.66	-0.83	3.33	4.17	1.39	0.56	-0.27	2.51	-0.82	1.96

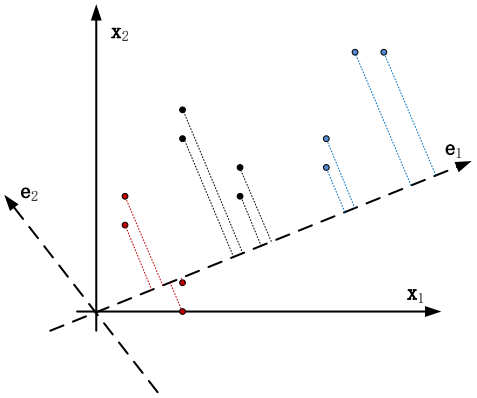

图 5.8 例 5.3 的样本与投影方向

选择最大特征值对应的特征矢量$\mathbf{e}_1$作为投影方向可以将所有样本降为 1 维特征。从图 5.8 可以看出样本在$\mathbf{e}_1$方向上的投影具有最大的可分性，对于 3 类问题只需要选择两个阈值即可对其分类。■

基于 Fisher 准则可分性分析的相关问题

与主成分分析不同，基于 Fisher 准则的可分性分析是通过计算矩阵$\mathbf{S}_w^{-1}\mathbf{S}_b$的特征值和特征矢量得到最优降维坐标投影方向的。但是$\mathbf{S}_w^{-1}\mathbf{S}_b$一般来说不是对称矩阵，虽然可以证明它的特征值和特征矢量都会是实数，但特征矢量之间并不具有正交性，因此降维之后新的坐标系不再是直角坐标系。非直角坐标系下仍然可以表示每一个特征矢量，只是特征之间具有一定的相关性，对于后续的分类器设计和学习并不会带来不利的影响。

在此之前我们总是假定矩阵 $\mathbf{S}_w$ 是一个可逆矩阵，然后计算 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值分解。在实际问题中 $\mathbf{S}_w$ 存在奇异矩阵的可能性，逆矩阵可能不存在，一般来说当训练样本数足够多时（ $n > d$ ），可以保证 $\mathbf{S}_w$ 是非奇异的，而当样本数 n 小于原始特征维数 d 时，$\mathbf{S}_w$ 为奇异矩阵。

基于 Fisher 准则的可分性分析算法中需要确定一个参数 d'，即降维之后的特征矢量维数。从前面的分析可以看出，矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 特征值的大小描述了向相应特征矢量方向投影的可分性。可以证明，对于 c 个类别的样本集来说，矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 至多只存在 $c-1$ 个特征值大于等于 0，其它的 $d-c+1$ 个特征值均为 0。由此可以看出对于 d' 的选择是有一定限制的，当类别数较大时我们可以根据情况选择 $d' < c-1$，但当类别数较少时只能选择 d' 为 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 非 0 特征值的个数。

基于 Fisher 准则的可分性分析有时也被用来作为一种线性分类器的学习方法。因为一旦找到了可以使得两个类别样本可分性最强的投影方向矢量 $\mathbf{w}$，那么就可以将所有的训练样本和待识别样本变换为 1 维特征。在一维空间中区分两类样本最简单的方法是设定一个合适的阈值 $-w_0$，如果样本 $\mathbf{x}$ 在 $\mathbf{w}$ 方向上的投影大于阈值则判别为 ω_1 类，否则为 ω_2 类，即：

$$\mathbf{w}^t\mathbf{x}\begin{cases}\geq -w_0, & \mathbf{x}\in\omega_1\\ < -w_0, & \mathbf{x}\in\omega_2\end{cases}$$

定义 $g(\mathbf{x})=\mathbf{w}^t\mathbf{x}+w_0$，则有：

$$g(\mathbf{x})=\mathbf{w}^t\mathbf{x}+w_0\begin{cases}\geq 0, & \mathbf{x}\in\omega_1\\ < 0, & \mathbf{x}\in\omega_2\end{cases} \qquad (5.52)$$

显然，这种判别方法同第 4 章公式（4.4）是一致的，$g(\mathbf{x})$ 即为线性判别函数。至于合适的阈值 $-w_0$ 可以采用第 7 章中将要介绍的贝叶斯决策理论来确定。

函数名称：FDA
参数：X--样本矩阵（n×d 矩阵），T--样本所属类别（n×1 矩阵）（1,2,…）
返回值：E--基矢量矩阵（d×d'矩阵，每列一个矢量）
函数功能：基于 Fisher 准则的可分性分析

```
function E = FDA( X, T )

d = size(X,2);
c = length(unique(T));

cmu = mean(X);
mu = zeros(c,d);
Sb = zeros(d,d);
Sw = zeros(d,d);
for i = 1:c                                          %计算类内、类间散布矩阵
    id = find( T == i );
    mu(i,:) = mean(X(id,:));
    ni = length(id);
```

```
    Sb = Sb + ni * ( mu(i,:) - cmu )' * ( mu(i,:) - cmu );
    Sw = Sw + ( X(id,:) - repmat(mu(i,:),ni,1) )' * ( X(id,:) - repmat(mu(i,:),ni,1) );
end
[V,L] = eigs( Sb, Sw, c-1 );                    %计算c-1个投影矢量
E = V(:,1:c-1);
```

函数名称：FDADR

参数：X--样本矩阵（n×d 矩阵），E--基矢量矩阵（d×d'矩阵）

返回值：Y--降维之后的样本矩阵（n×d'矩阵）

函数功能：基于 Fisher 准则的可分性分析降维

```
function Y = FDADR(X,E)

Y = X*E;
```

小结

人们在面对复杂的模式识别问题时，往往无法通过主观观察确定哪些特征对分类来说是必须的，哪些特征中不包含类别之间的可分性信息。大多数的识别系统设计者会生成尽量多的原始特征，然后使用特征选择与提取的方法来找出其中蕴含的类别可分性信息，降低识别特征的维数。近年来在实际应用需求的推动下，对于特征降维方法的研究受到了广泛的重视，本章只是介绍了这方面最基本的一些知识和方法。

特征选择希望从原始特征中挑选出一组包含最多类别可分性信息的特征，从当前的研究情况来看这实际上是非常困难的。问题的难度首先体现在到目前为止，还没有一种能够准确评价一组特征对某个分类问题有效程度的方法。使用这组特征设计和学习一个分类器，以分类器的识别准确率来评价特征的有效性可能是一种最有效的手段，采用这种方法仍然要受到分类器设计过程中选择的不同识别方法、模型参数和学习算法的影响。特征选择另一个难题是如何找到这组“最优”的特征。由于特征之间并不独立，存在着相关性，因此无法单独评价每一个特征，需要对不同的特征组合进行评价，目前还没有一个有效的算法能够在多项式时间内完成这项工作。分支定界法是到目前为止唯一能够在一定程度上减少计算量的最优搜索算法，但它的最优性需要可分性准则的单调性来保证，在模式识别系统设计过程中转而寻找次优的特征组合是一种更加实际可行方案。

主成分分析的方法产生于 20 世纪初，它的目的是要由原始变量的线性组合产生出一组互不相关的新的变量，类似的方法在信号处理领域称为 Karhunen-Loève 变换。由于计算简单、适用性强，主成分分析已经被成功地用于解决各种不同的分类问题。例如，在人脸识别中一种常用的方法是将人脸图像经过预处理后，以每个像素点的灰度值为特征构成识别特征矢量，对于一个128×128的人脸图像来说特征维数就会达到 16384 维，以大量不同的人脸图像作为训练样本，采用主成分分析的方法寻找这些特征中若干主要分量实现特征的降维，这些主要的分量也被称为“特征人脸”(Eigenface)；在自然语言处理领域的文本分类中也会面临同样的问题，如果以文本中不同词出现的频度为特征，就会得到一个上万维的特征

矢量，将所有文本的特征矢量排成一个文本矩阵，计算文本矩阵的奇异值分解(Singular Value Decomposition, SVD)，保留最大的若干个奇异值对应的左奇异向量，由这些奇异矢量可以将原始的文本词频特征矢量降维，这种方法一般被称作潜在语义分析(Latent Semantic Analysis, LSA)，可以证明矩阵的奇异值是相应特征值的非负平方根，而左奇异矢量则是相应协方差矩阵的特征矢量。

主成分分析去除了特征之间的相关性，以此实现了特征维度的降低。在使用过程中需要注意的是，主成分分析是一种无监督学习方法，小的特征值只是说明相应维度上样本分布的方差较小，并不代表它对分类的作用小。基于 Fisher 准则的可分性分析以保留最多的类别可分性信息为目标来降低特征维度，计算过程中需要使用训练样本的类别属性信息，是一种有监督的学习方法。这里需要注意的是基于 Fisher 准则的可分性分析只能够得到不超过 $c-1$ 维的降维识别特征，当类别数量较少时，过少的特征可能并不足以区分各个类别。这时需要考虑采用一些改进的方法，例如可以首先使用主成分分析的方法寻找到样本分布的主分量，然后不以相应特征值的大小来决定保留哪些分量，而是使用 Fisher 准则来评估各个分量上的类别可分性。由于主成分分析去除了特征之间的相关性，因此可分性判据最大的一组分量可以认为是最优的特征组合。

除了上述两种方法之外，近年来很多统计学中发展起来的成分分析方法也被用于识别特征的降维。主成分分析可以使得降维之后的特征之间是不相关的，但并不能保证是相互独立的，独立成分分析（Independent Component Analysis, ICA）就是以追求变换之后特征之间具有独立性为目的实现的特征降维；其它常用的方法还包括：多维尺度变换（Multidimensional Scaling, MDS）、典型相关分析（Canonical Correlation Analysis, CCA）、偏最小二乘（Partial Least Square, PLS）等等。

特征提取并不局限于上述这些线性变换的方法，近年来也提出了很多非线性的特征降维方法，这些方法一般被称作“流形学习”（Manifold Learning）。非线性特征提取方法中有一类是以线性方法为基础，结合第 6 章中将要介绍的“核方法”实现的，如核主成分分析、核线性判别分析等等，一般来说所有的线性特征提取方法都可以通过引入核函数变成非线性方法；另一类方法是利用非线性流形在局部可以用线性流形近似的特点实现的非线性特征提取，如 Isomap 和 Locally Linear Embedding（LLE）。

习题

1. 证明公式（5.3）和（5.4）的等价性。
2. 证明当采用欧氏距离度量时，公式（5.6）和公式（5.7a）、（5.7b）是等价的。
3. 证明总体散布矩阵为类内散布矩阵和类间散布矩阵之和，即公式（5.11）成立。
4. 验证类内、类间均方距离 J_{msd} 、 J_{bsd} 以及类别可分性准则 J_1 和 J_3 满足（5.18）式所定义的单调性。
5. 分别给出两种方式增 l-减 r 算法的计算复杂度，一种方式是由空集 Φ 开始增加（ $l>r$ ），

另一种方式是由全集 $\mathcal{X}$ 开始减少（$l < r$）。

6. 使用主成分分析的方法将下列 2 维样本降维为 1 维特征，并求出降维之后每个样本的特征值，在二维空间中画出样本和投影坐标轴：

$$\mathrm{x}_1 = (23,22)^t，\ \mathrm{x}_2 = (22,23)^t，\ \mathrm{x}_3 = (0,0)^t，\ \mathrm{x}_4 = (10,11)^t，\ \mathrm{x}_5 = (11,10)^t$$

7. 证明样本集协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值均为不小于 0 的实数，并且特征矢量均为实矢量。
8. 证明样本集协方差矩阵 $\boldsymbol{\Sigma}$ 的任意两个不同特征值对应的特征矢量之间是正交的。
9. 当原始特征维数 d 比较大时，求取协方差矩阵 $\boldsymbol{\Sigma}$ 全部的特征值和特征矢量的计算量比较大，改进 PCA 的 Matlab 代码，逐个计算最大的特征值和特征矢量，直到累加和达到设定的比例 ratio 为止。
10. 推导 c 个类别情况下基于 Fisher 准则的可分性分析过程。
11. 有两类样本集：

$$\omega_1 = \left\{(0,0,0)^t,(1,0,0)^t,(1,0,1)^t,(1,1,0)^t\right\}$$

$$\omega_2 = \left\{(0,0,1)^t,(0,1,0)^t,(0,1,1)^t,(1,1,1)^t\right\}$$

请使用 Fisher 线性判别方法将 3 维特征降为 1 维，并在 1 维空间中构建分类器区分两个类别。

12. 证明在基于 Fisher 准则的可分性分析中，矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值均为大于等于 0 的实数，所有的特征矢量均为实矢量。
13. 证明由 c 个类别的样本计算的矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 最多只有 $c-1$ 特征值大于 0。
14. 证明当样本数 n 小于原始特征维数 d 时，$\mathbf{S}_w$ 为奇异矩阵。