Linear Discriminant Analysis

Roger Jang
MATLAB code & examples
2010/02/11

線性識別分析

線性識別分析（linear discriminant analysis，簡稱 LDA）的主要概念乃是希望能找出最適合的投影方向，使得一群已經事先分類完成的資料點在投影到低維度空間後，屬於同一類別的資料點能盡量集中，而不同類別之間的資料點能盡量分開。

假設我們共有 n 個資料點，每一點的維度都是 m，這些資料點可以使用集合表示成 $\{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_m\}$，或者使用一個 $m\times n$ 的矩陣 A 表示成：

$$\mathbf{A}=\begin{bmatrix} | & & | & & | \\ \mathbf{a}_1 & \cdots & \mathbf{a}_i & \cdots & \mathbf{a}_n \\ | & & | & & | \end{bmatrix}$$

這些資料點的平均值可以寫成 $\mu=\frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i$。假設這 m 筆資料分別屬於 c 種類別，則資料矩陣 A 亦可表示成：

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_1 & \cdots & \mathbf{A}_k & \cdots & \mathbf{A}_c\end{bmatrix}$$

其中，$\mathbf{A}_k$ 表示第 k 個類別的資料矩陣，$k=1,2,\ldots,c$。假設第 k 類資料的資料個數為 m_k 個，那麼我們便可以得到下列兩式：

$$m=\sum_{k=1}^{c}m_k$$

$$m\mu=\sum_{k=1}^{c}m_k\mu_k$$

首先我們將這些資料平移到零點，並計算出這些資料點在單位向量 $\mathbf{d}$ 的投影平方和，如下：

$$
\begin{aligned}
\textit{total squared projection} &= \sum_{i=1}^{n}\left[(\mathbf{a}_i-\mu)^T\mathbf{d}\right]^2 \\
&= \sum_{i=1}^{n}\left[\mathbf{d}^T(\mathbf{a}_i-\mu)\right]\left[(\mathbf{a}_i-\mu)^T\mathbf{d}\right] \\
&= \sum_{i=1}^{n}\mathbf{d}^T(\mathbf{a}_i-\mu)(\mathbf{a}_i-\mu)^T\mathbf{d} \\
&= \mathbf{d}^T\left[\sum_{i=1}^{n}(\mathbf{a}_i-\mu)(\mathbf{a}_i-\mu)^T\right]\mathbf{d} \\
&= \mathbf{d}^T\mathbf{T}\mathbf{d}
\end{aligned}
$$

在上式中，我們可用$\mathbf{d}^T\mathbf{T}\mathbf{d}$來表示將資料點投影在單位向量$\mathbf{d}$方向的投影平方和，因此$T$稱為全部資料點的散佈矩陣（total scatter matrix），可表示如下：

$$
\begin{aligned}
\mathbf{T} &= \sum_{i=1}^{n}(\mathbf{a}_i-\mu)(\mathbf{a}_i-\mu)^T \\
&= \sum_{i=1}^{n}\left(\mathbf{a}_i\mathbf{a}_i^T-\mathbf{a}_i\mu^T-\mu\,\mathbf{a}_i^T+\mu\;\mu^T\right) \\
&= \sum_{i=1}^{n}\mathbf{a}_i\mathbf{a}_i^T-\sum_{i=1}^{n}\mathbf{a}_i\mu^T-\sum_{i=1}^{n}\mu\,\mathbf{a}_i^T+\sum_{i=1}^{n}\mu\;\mu^T \\
&= \sum_{i=1}^{n}\mathbf{a}_i\mathbf{a}_i^T-\left[\sum_{i=1}^{n}\mathbf{a}_i\right]\mu^T-\mu\left[\sum_{i=1}^{n}\mathbf{a}_i^T\right]+\mathbf{n}\,\mu\;\mu^T \\
&= \sum_{i=1}^{n}\mathbf{a}_i\mathbf{a}_i^T-n\,\mu\;\mu^T \\
&= \mathbf{A}\mathbf{A}^T-n\,\mu\;\mu^T
\end{aligned}
$$

此散佈矩陣滿足下列恆等式：

$$\mathbf{T}=\mathbf{B}+\mathbf{W}$$

其中$\mathbf{B}$表示「類別間散佈矩陣」（between-class scatter matrix），代表若將每個類別的資料視為一個單獨的資料點（由此類別的平均向量來代表）所的到的散佈矩陣。而$\mathbf{W}$表示「類別內散佈矩陣」（within-class scatter matrix），代表若將每個類別的資料視為一個獨立資料集合所得到的散佈矩陣。欲證明上述恆等式，我們可將$\mathbf{B}$以及$\mathbf{W}$的定義化簡如下：

$$
\begin{aligned}
\mathbf{B} &= \sum_{k=1}^{c}m_k(\mu_k-\mu)(\mu_k-\mu)^T \\
&= \sum_{k=1}^{c}m_k\mu_k\mu_k^T-n\,\mu\;\mu^T \\
\mathbf{W} &= \sum_{k=1}^{c}\mathbf{W}_i \\
&= \sum_{k=1}^{c}(\mathbf{A}_k\mathbf{A}_k^T-m_k\mu_k{\mu_k}^T) \\
&= \mathbf{A}\mathbf{A}^T-\sum_{k=1}^{c}m_k\mu_k{\mu_k}^T
\end{aligned}
$$

由此可看出上述恆等式是一定成立的。

根據上述恆等式，我們可以將投影量的平方和拆成分成兩部分：

$$\mathbf{d}^T\mathbf{T}\mathbf{d}=\mathbf{d}^T\mathbf{B}\mathbf{d}+\mathbf{d}^T\mathbf{W}\mathbf{d}$$

由直觀的角度來看，線性識別分析的目的，乃是希望找到某個單位投影向量$\mathbf{d}$，以便在投影後，盡量增大類別間投影量平方和$\mathbf{d}^T\mathbf{B}\mathbf{d}$，並盡量縮小類別內投影量平方和$\mathbf{d}^T\mathbf{W}\mathbf{d}$，換句話說，亦即希望資料在投影後，隸屬於同一類別的資料點能盡量集中，而類別與類別之間能盡量散開。

雖然我們已經可以將投影量的平方和拆成兩部分，也知道選取投影向量$\mathbf{d}$的「大概」標準，但是在實作上，我們還是要定義一個目標函數，經由對此目標函數的最佳化，來求取最佳的投影向量$\mathbf{d}$。最直覺的目標函數可以定義如下：

$$J(\mathbf{d})=\frac{\mathbf{d}^T\mathbf{B}\mathbf{d}}{\mathbf{d}^T\mathbf{W}\mathbf{d}}$$

根據上面的條件，我們可以把識別分析的步驟歸納成下面三點：

1. 尋找能得到最大J值的單位向量 $\mathbf{d}_1$。
2. 在符合$\mathbf{d}_2^T\mathbf{W}\mathbf{d}_1=0$的條件下，尋找能得到最大$J$的單位向量$\mathbf{d}_2$。
3. 在符合$\mathbf{d}_k^T\mathbf{W}\mathbf{d}_i=0$的條件下，尋找能得到最大$J$的單位向量$\mathbf{d}_k$，其中$1\le i<k$。

我們將$\{\mathbf{d}_i\}$稱之為最能識別（discriminant）該組資料的識別向量（discriminant vectors）。值得一提的是，$\{\mathbf{d}_i\}$在這裡並未限定是必須兩兩互相垂直的向量集合，然而這樣的彈性卻使我們不易找出適當的目標函數（object function）來執行上述演算法。因此 Duchene 及 Leclercq 在 1988 年提出了新的觀點，他們根據上述演算法在第三步驟加入新的條件：$\mathbf{d}_k^T\mathbf{d}_i=0$且$\mathbf{d}_k^T\mathbf{W}\mathbf{d}_k=1$，並透過實驗說明新的演算法優於傳統的識別分析方法 [8]，茲說明如下。

假設目前我們希望尋找第 k 個識別向量$\mathbf{d}_k$，根據 Duchene 及 Leclercq 演算法中新增的條件，我們可以把識別分析的目標從「將J最大化」 簡化成「將$\mathbf{d}_k^T B\mathbf{d}_k$最大化」，並且滿足$\mathbf{d}_k^T\mathbf{d}_i=0$且$\mathbf{d}_k^T\mathbf{W}\mathbf{d}_k=1$。根據上述最佳化的目標函數以及相關的限制條件，我們可以藉由 Lagrange multiplier 定義出新的目標函數如下（為簡化數學符號起見，我們使用$\mathbf{d}$來取代$\mathbf{d}_k$）：

$$\begin{aligned}J(\mathbf{d})&=\mathbf{d}^T\mathbf{B}\mathbf{d}-\lambda(\mathbf{d}^T\mathbf{W}\mathbf{d}-1)-\sum_{i=1}^{k-1}\rho_i\mathbf{d}^T\mathbf{d}_i\\&=\mathbf{d}^T\mathbf{B}\mathbf{d}-\lambda(\mathbf{d}^T\mathbf{W}\mathbf{d}-1)-\mathbf{d}^T\mathbf{D}\rho\end{aligned}$$

其中，$\rho=[\rho_1,\rho_2,\ldots,\rho_{k-1}]^T$，而$\mathbf{D}$包含 k-1 行，每一直行均由$\mathbf{d}_i$組成：

$$\mathbf{D}=\begin{bmatrix} | & | & & | \\ \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_{k-1} \\ | & | & & | \end{bmatrix}$$

為了求$J(\mathbf{d})$的極值，我們計算$J(\mathbf{d})$的梯度（gradient）如下：

$$2\mathbf{B}\mathbf{d}-2\lambda\mathbf{W}\mathbf{d}-\mathbf{D}\rho=0$$
$$\Rightarrow\quad \mathbf{D}\rho=2\mathbf{B}\mathbf{d}-2\lambda\mathbf{W}\mathbf{d} \qquad (7\text{-}2\text{-}1\text{-}\mathrm{A}\text{-}2)$$

若要消去λ，我們可將方程式（7-2-1-A-2）乘上$\mathbf{D}^{\mathrm{T}}\mathbf{W}^{-1}$，便可以得到：

$$\mathbf{D}^T\mathbf{W}^{-1}\mathbf{D}\rho=2\mathbf{D}^T\mathbf{W}^{-1}\mathbf{B}\mathbf{d}$$
$$\Rightarrow\quad \rho=2\left(\mathbf{D}^T\mathbf{W}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{W}^{-1}\mathbf{B}\mathbf{d} \qquad (7\text{-}2\text{-}1\text{-}\mathrm{A}\text{-}3)$$

（上式的簡化用到了原問題的限制條件$\mathbf{D}^{\mathrm{T}}\mathbf{d}=\mathbf{0}$。）

將方程式（7-2-1-A-3）代入（7-2-1-A-2）中，我們可以得到：

$$2\mathbf{D}\left(\mathbf{D}^T\mathbf{W}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{W}^{-1}\mathbf{B}\mathbf{d}=2\mathbf{B}\mathbf{d}-2\lambda\mathbf{W}\mathbf{d}$$

再將上式乘上 $0.5\mathbf{W}^{-1}$，我們便可以得到：

$$\mathbf{W}^{-1}\mathbf{D}\left(\mathbf{D}^T\mathbf{W}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{W}^{-1}\mathbf{B}\mathbf{d}=\mathbf{W}^{-1}\mathbf{B}\mathbf{d}-\lambda\mathbf{d}$$
$$\Rightarrow\quad \left(I-\mathbf{W}^{-1}\mathbf{D}\left(\mathbf{D}^T\mathbf{W}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}^T\right)\mathbf{W}^{-1}\mathbf{B}\mathbf{d}=\lambda\mathbf{d} \qquad (7\text{-}2\text{-}1\text{-}\mathrm{A}\text{-}4)$$

由方程式（7-2-1-A-4），我們可以得出結論，第k個識別向量$\mathbf{d}$即為矩陣$\mathbf{Q}=\left(I-\mathbf{W}^{-1}\mathbf{D}\left(\mathbf{D}^T\mathbf{W}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}^T\right)\mathbf{W}^{-1}\mathbf{B}$最大特徵值所對應到的特徵向量。（為什麼？請想想看！）另外要注意的是，在計算$\mathbf{d}_1$時，$\mathbf{D}$是一個空矩陣，因此當$k=1$時，$\mathbf{Q}=\mathbf{W}^{-1}\mathbf{B}$。