LDA 推导

Soul Walker

2020 年 7 月 22 日

给定数据矩阵 $\boldsymbol{X} = (\boldsymbol{x}_1; \boldsymbol{x}_2; \cdots; \boldsymbol{x}_n) \in R^{n\times m}$，标签矩阵 $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)$，$X_i(i = 1, 2, \cdots, k)$ 为属于第 i 类的样本集合，$n_i(i = 1, 2, \cdots, k)$ 为属于第 i 类的样本个数，训练数据集

$$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$$

其中，$y_i \in \{C_1, C_2, \cdots, C_k\}$ 为实例的类别。定义 $\boldsymbol{\mu}_i(i = 1, 2, \cdots, k)$ 为第 i 类样本的均值向量，定义样本整体均值 $\boldsymbol{\mu}$，$\boldsymbol{\Sigma}_i(i = 1, 2, \cdots, k)$ 为第 i 类样本的协方差矩阵，并且我们希望将数据降到 d 维。

0.1 散度矩阵

首先我们定义变换

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^T\boldsymbol{x}_i \quad \in R^d(d \leq k-1) \tag{0.1}$$

其中 $\boldsymbol{w} \in R^{n\times d}$，对于 $d < k-1$ 的问题有如下解释，因为 S_b（后文做推导）中每个 $\boldsymbol{\mu}_i - \boldsymbol{\mu}$ 的秩 (rank) 为 1，因此相加后的协方差矩阵秩小于等于 k，但是我们知道前 $k-1$ 个 $\boldsymbol{\mu}_i$ 之后，最后一个 $\boldsymbol{\mu}_k$ 可以由前 $k-1$ 个 $\boldsymbol{\mu}_i$ 线性表示，因此 $d \leq k-1$ 成立

我们的目标是使得类内距离小，类间距离大。

对于类内（协方差矩阵）:

$$\begin{aligned}
&\sum_{i\in C_k}(\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{w}^T\boldsymbol{\mu}_k)(\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{w}^T\boldsymbol{\mu}_k)^T \\
=&\boldsymbol{w}^T\sum_{i\in C_k}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{w}
\end{aligned} \tag{0.2}$$

其中我们定义 $S_k = \sum_{i\in C_k}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T$，上式可以做如下变换

$$
\begin{aligned}
&\boldsymbol{w}^T \sum_{i\in C_k}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{w} \\
=&tr(\boldsymbol{w}^T S_k \boldsymbol{w}) \\
=&tr(\boldsymbol{w}^T \sum_{i=1}^{k} S_i \boldsymbol{w})
\end{aligned} \tag{0.3}
$$

我们令 $S_w = \sum_{i=1}^{k}$，这里称 S_w 为类内散度矩阵，那么上式最终变换得

$$
\sum_{i\in C_k}(\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{w}^T\boldsymbol{\mu}_k)(\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{w}^T\boldsymbol{\mu}_k)^T = tr(\boldsymbol{w}^T S_w \boldsymbol{w}) \tag{0.4}
$$

对于类间（协方差矩阵）:

$$
\begin{aligned}
d_i =&(\boldsymbol{w}^T\boldsymbol{\mu}_i - \boldsymbol{w}^T\boldsymbol{\mu})^T(\boldsymbol{w}^T\boldsymbol{\mu}_i - \boldsymbol{w}^T\boldsymbol{\mu}) \\
=&(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T\boldsymbol{w}\boldsymbol{w}^T(\boldsymbol{\mu}_i - \boldsymbol{\mu})
\end{aligned} \tag{0.5}
$$

我们令 $\boldsymbol{M} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k)$，$\bar{\boldsymbol{M}} = (\boldsymbol{\mu}, \cdots, \boldsymbol{\mu})$

$$
\begin{aligned}
&\sum_{i=1}^{k} d_i \\
=&\sum_{i=1}^{k}(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T\boldsymbol{w}\boldsymbol{w}^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}) \\
=&tr((\boldsymbol{M} - \bar{\boldsymbol{M}})^T\boldsymbol{w}\boldsymbol{w}^T(\boldsymbol{M} - \bar{\boldsymbol{M}})) \\
=&tr(\boldsymbol{w}^T(\boldsymbol{M} - \bar{\boldsymbol{M}})(\boldsymbol{M} - \bar{\boldsymbol{M}})^T\boldsymbol{w})
\end{aligned} \tag{0.6}
$$

这里我们令 $S_b = (\boldsymbol{M} - \bar{\boldsymbol{M}})(\boldsymbol{M} - \bar{\boldsymbol{M}})^T$，这里 S_b 称为类间散度矩阵，那么上式最终变换得

$$
\sum_{i=1}^{k} d_i = tr(\boldsymbol{w}^T S_b \boldsymbol{w}) \tag{0.7}
$$

0.2 目标及求解

我们的目标是使得类内距离小，类间距离大，即最优化问题:

$$
\max_{\boldsymbol{w}} \frac{tr(\boldsymbol{w}^T S_b \boldsymbol{w})}{tr(\boldsymbol{w}^T S_w \boldsymbol{w})} \tag{0.8}
$$

那么我们的约束最优化问题为

$$\begin{cases} \max\limits_{\boldsymbol{w}} tr(\boldsymbol{w}^T S_b \boldsymbol{w}) \\ s.t. \quad tr(\boldsymbol{w}^T S_w \boldsymbol{w}) = d \end{cases} \tag{0.9}$$

对上式求解可以使用拉格朗日乘子法如下

$$L(\boldsymbol{w}, \boldsymbol{\Lambda}) = tr(\boldsymbol{w}^T S_b \boldsymbol{w}) - tr(\boldsymbol{w}^T S_w \boldsymbol{w} \boldsymbol{\Lambda}) \tag{0.10}$$

其中

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \tag{0.11}$$

对上式求偏导得

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}, \boldsymbol{\Lambda})}{\partial \boldsymbol{w}} &= 0 \\ S_b \boldsymbol{w} &= S_w \boldsymbol{w} \boldsymbol{\Lambda} \\ S_w^{-1} S_b \boldsymbol{w} &= S_w^{-1} S_w \boldsymbol{w} \boldsymbol{\Lambda} \\ S_w^{-1} S_b \boldsymbol{w} &= \boldsymbol{w} \boldsymbol{\Lambda} \end{aligned} \tag{0.12}$$

那么 $\boldsymbol{w}$ 的求解由 $S_w^{-1} S_b$ 的最大的 k 个特征向量组成。
这里有一个小问题：对于 S_w 是否可逆的讨论，如果不可逆的话可进行：

1 . $S_w + I_d \cdot 0.001$(or else);

2 . 先做 PCA。