AMSC660 Final Report

Department of Chemistry and Biochemistry, University of Maryland

Kecai Xuan

kxuan@umd.edu

December 10, 2024

1 Linear Discriminant Analysis

1.1 Rank of $\mathrm{S_b}$

Consider the vector set $\{m_1 - m, m_2 - m, \cdots m_c - m\}$, we have the following relation:

$$\sum_{i=1}^{c} n_i \left(m_i - m\right) = \left[\sum_{i=1}^{c} n_i m_i\right] - nm = nm - nm = 0$$

which means that there exist the nonzero linear combination parameters such that the sum of the vectors is zero. Therefore, the vector set is linearly dependent.

Since S_b is defined as

$$S_b = \sum_{i=1}^{c} n_i \left(m_i - m\right) \left(m_i - m\right)^\top$$

a weighted sum of these rank-1 matrices, it follows that the image (or output) of the linear transformation defined by S_b lies within the space spanned by the vector set $\{m_1 - m, m_2 - m, \cdots m_c - m\}$. That is because when applying S_b to any vector x, it gives the linear combination of the vectors in the set:

$$S_b x = \sum_{i=1}^{c} n_i \left(m_i - m\right) \left(m_i - m\right)^\top x$$

where $\left(m_i - m\right)^\top x$ is a scalar.

The rank of matrix S_b is equal to the dimension of its image, which is the space spanned by the vector set $\{m_1 - m, m_2 - m, \cdots m_c - m\}$ that is linearly dependent. Therefore, the rank of S_b is at most $c - 1$.

1.2 Gradient of $\mathbf{J(w)}$

The gradient of $J(w)$ is given by

$$\begin{aligned}
\nabla J(w) &= \nabla \left(\frac{w^\top S_b w}{w^\top S_w w}\right) \\
&= \frac{2}{(w^\top S_w w)^2} \left[(w^\top S_w w) S_b w - (w^\top S_b w) S_w w\right]
\end{aligned}$$

Let it be zero, we have $(w^\top S_w w)S_b w - (w^\top S_b w)S_w w = 0$, where $w^\top S_w w, w^\top S_b w$ are scalars. So the only way to get the equation satisfied is to have

$$S_b w = \lambda S_w w, \quad \lambda = \frac{w^\top S_b w}{w^\top S_w w}$$

1.3 Cholesky Decomposition

It is obvious that S_w is a SPD matrix, which can be decomposed as $S_w = LL^\top$. Define $y = L^\top w, w = L^{-\top} y$, then the equation $S_b w = \lambda S_w w$ can be rewritten as

$$S_b L^{-\top} y = \lambda L y \quad \Rightarrow \quad L^{-1} S_b L^{-\top} y = \lambda y$$

which takes the form of $Ay = \lambda y$ where $A = L^{-1} S_b L^{-\top}$ is SPD.

1.4 Comparison between LDA and PCA on MINST

The projection of the data onto the first two linear discriminants and the first two principal components are shown in the following figure. It is clear that the linear discriminants are able to separate the data better than the principal components.