LECTURE 10: Linear Discriminant Analysis

- Linear Discriminant Analysis, two classes
- Linear Discriminant Analysis, C classes
- LDA vs. PCA example
- Limitations of LDA
- Variants of LDA
- Other dimensionality reduction methods
The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible.

- Assume we have a set of D-dimensional samples \( \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \), \( N_1 \) of which belong to class \( \omega_1 \), and \( N_2 \) to class \( \omega_2 \). We seek to obtain a scalar \( y \) by projecting the samples \( x \) onto a line

\[
y = w^T x
\]

- Of all the possible lines we would like to select the one that maximizes the separability of the scalars.
  - This is illustrated for the two-dimensional case in the following figures.
In order to find a good projection vector, we need to define a measure of separation between the projections.

- The mean vector of each class in $x$ and $y$ feature space is
  \[
  \mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x \quad \text{and} \quad \tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \omega_i} y = \frac{1}{N_i} \sum_{x \in \omega_i} w^T x = w^T \mu_i
  \]

- We could then choose the distance between the projected means as our objective function
  \[
  J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |w^T (\mu_1 - \mu_2)|
  \]

- However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within the classes.

---

This axis yields better class separability

This axis has a larger distance between means
Linear Discriminant Analysis, two-classes (3)

- The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter
  - For each class we define the scatter, an equivalent of the variance, as
    \[ \tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu})^2 \]
    - where the quantity \( (\tilde{s}_1^2 + \tilde{s}_2^2) \) is called the within-class scatter of the projected examples
  - The Fisher linear discriminant is defined as the linear function \( w^T x \) that maximizes the criterion function
    \[ J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \]
    - Therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible
Linear Discriminant Analysis, two-classes (4)

- In order to find the optimum projection \( w^* \), we need to express \( J(w) \) as an explicit function of \( w \).
- We define a measure of the scatter in multivariate feature space \( x \), which are scatter matrices

\[
S_1 = \sum_{x \in \omega_1} (x - \mu)(x - \mu)^T
\]

\[
S_1 + S_2 = S_W
\]

- where \( S_W \) is called the **within-class scatter matrix**
- The scatter of the projection \( y \) can then be expressed as a function of the scatter matrix in feature space \( x \)

\[
\tilde{s}_1^2 = \sum_{y \in \omega_1} (y - \tilde{\mu})^2 = \sum_{x \in \omega_1} (w^T x - w^T \mu)^2 = \sum_{x \in \omega_1} w^T (x - \mu)(x - \mu)^T w = w^T S_1 w
\]

\[
\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w
\]

- Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space

\[
(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = w^T S_B w
\]

- The matrix \( S_B \) is called the **between-class scatter**. Note that, since \( S_B \) is the outer product of two vectors, its rank is at most one
- We can finally express the Fisher criterion in terms of \( S_W \) and \( S_B \) as

\[
J(w) = \frac{w^T S_B w}{w^T S_W w}
\]
Linear Discriminant Analysis, two-classes (5)

- To find the maximum of $J(w)$ we derive and equate to zero

$$\frac{d}{dw}[J(w)] = \frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right] = 0 \Rightarrow$$

$$\Rightarrow \left[w^T S_w w\right]\frac{d\left[w^T S_B w\right]}{dw} - \left[w^T S_B w\right]\frac{d\left[w^T S_w w\right]}{dw} = 0 \Rightarrow$$

$$\Rightarrow \left[w^T S_w w\right]2S_B w - \left[w^T S_B w\right]2S_w w = 0$$

- Dividing by $w^T S_w w$

$$\frac{\left[w^T S_w w\right]}{w^T S_w w} S_B w - \frac{\left[w^T S_B w\right]}{w^T S_w w} S_w w = 0 \Rightarrow$$

$$\Rightarrow S_B w - JS_w w = 0 \Rightarrow$$

$$\Rightarrow S_w^{-1} S_B w - Jw = 0$$

- Solving the generalized eigenvalue problem ($S_w^{-1} S_B w = Jw$) yields

$$w^* = \arg \max_w \left\{\frac{w^T S_B w}{w^T S_w w}\right\} = S_w^{-1}(\mu_1 - \mu_2)$$

- This is know as Fisher's Linear Discriminant (1936), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension
LDA example

- Compute the Linear Discriminant projection for the following two-dimensional dataset
  - $X_1 = \{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
  - $X_2 = \{(9,10),(6,8),(9,5),(8,7),(10,8)\}$

- **SOLUTION (by hand)**
  - The class statistics are:
    - $S_1 = \begin{bmatrix} 0.80 & -0.40 \\ -0.40 & 2.60 \end{bmatrix}$; $S_2 = \begin{bmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{bmatrix}$
    - $\mu_1 = \begin{bmatrix} 3.00 \\ 3.60 \end{bmatrix}$; $\mu_2 = \begin{bmatrix} 8.40 \\ 7.60 \end{bmatrix}$
  - The within- and between-class scatter are
    - $S_B = \begin{bmatrix} 29.16 & 21.60 \\ 21.60 & 16.00 \end{bmatrix}$; $S_W = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$
  - The LDA projection is then obtained as the solution of the generalized eigenvalue problem
    - $S_W^{-1}S_Bv = \lambda v$ \(\Rightarrow\) $|S_W^{-1}S_B - \lambda I| = 0$ \(\Rightarrow\) \(\begin{bmatrix} 11.89 - \lambda & 8.81 \\ 5.08 & 3.76 - \lambda \end{bmatrix}\) = 0 \(\Rightarrow\) $\lambda = 15.65$
    - \(\begin{bmatrix} 11.89 \\ 5.08 \end{bmatrix}v_1 = 15.65 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\) \(\Rightarrow\) \(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\) = \(\begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}\)
  - Or directly by
    - $w^* = S_W^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} -0.91 & -0.39 \end{bmatrix}^T$
Linear Discriminant Analysis, C-classes (1)

- Fisher’s LDA generalizes very gracefully for C-class problems
  - Instead of one projection $y$, we will now seek $(C-1)$ projections $[y_1, y_2, \ldots, y_{C-1}]$ by means of $(C-1)$ projection vectors $w_i$, which can be arranged by columns into a projection matrix $W=[w_1|w_2|\ldots|w_{C-1}]$:
    \[ y_i = w_i^T x \Rightarrow y = W^T x \]

- Derivation
  - The generalization of the within-class scatter is
    \[ S_W = \sum_{i=1}^{C} S_i \]
    where $S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$ and $\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$
  - The generalization for the between-class scatter is
    \[ S_B = \sum_{i=1}^{C} N_i(\mu - \mu_i)(\mu_i - \mu)^T \]
    where $\mu = \frac{1}{N} \sum_x x = \frac{1}{N} \sum_{x \in \omega_i} N_i \mu_i$
    - where $S_T = S_B + S_W$ is called the total scatter matrix
Similarly, we define the mean vector and scatter matrices for the projected samples as

\[
\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \omega_i} y \\
\tilde{\Sigma}_W = \sum_{i=1}^{C} \sum_{y \in \omega_i} (y - \tilde{\mu}_i)(y - \tilde{\mu}_i)^T \\
\tilde{\Sigma}_B = \sum_{i=1}^{C} N_i (\tilde{\mu}_i - \tilde{\mu})(\tilde{\mu}_i - \tilde{\mu})^T
\]

From our derivation for the two-class problem, we can write

\[
\tilde{\Sigma}_W = W^T \Sigma_W W \\
\tilde{\Sigma}_B = W^T \Sigma_B W
\]

Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scalar (it has C-1 dimensions), we then use the determinant of the scatter matrices to obtain a scalar objective function:

\[
J(W) = \frac{\det(\tilde{\Sigma}_B)}{\det(\tilde{\Sigma}_W)} = \frac{|W^T \Sigma_B W|}{|W^T \Sigma_W W|}
\]

And we will seek the projection matrix \( W^* \) that maximizes this ratio.
Linear Discriminant Analysis, C-classes (3)

- It can be shown that the optimal projection matrix $W^*$ is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem

$$W^* = [w_1^* | w_2^* | \cdots | w_{C-1}^*] = \arg\max \begin{vmatrix} W^T S_B W \end{vmatrix} \Rightarrow (S_B - \lambda_i S_W) w_i^* = 0$$

**NOTES**

- $S_B$ is the sum of $C$ matrices of rank one or less and the mean vectors are constrained by

$$\frac{1}{C} \sum_{i=1}^{C} \mu_i = \mu$$

  - Therefore, $S_B$ will be of rank (C-1) or less
  - This means that only (C-1) of the eigenvalues $\lambda_i$ will be non-zero

- The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $S_W^{-1} S_B$

- LDA can be derived as the Maximum Likelihood method for the case of normal class-conditional densities with equal covariance matrices