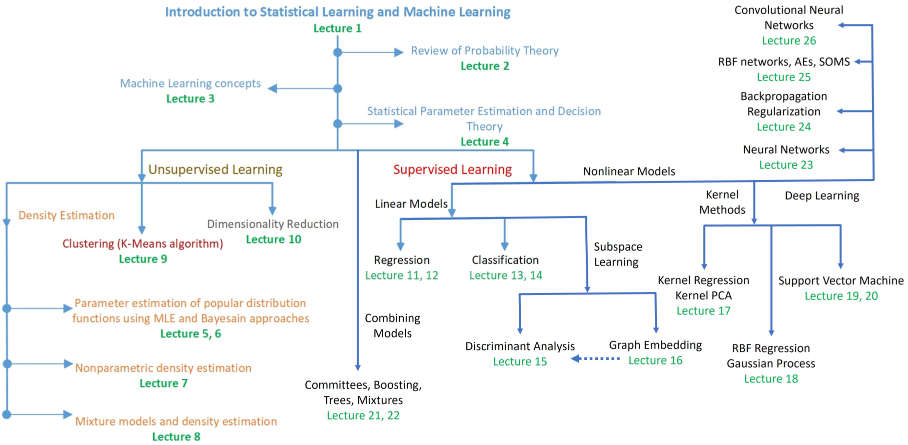
Statistical Learning and Machine Learning

Lecture 15 - Linear Discriminant Analysis

October 13, 2021

Course overview and where do we stand



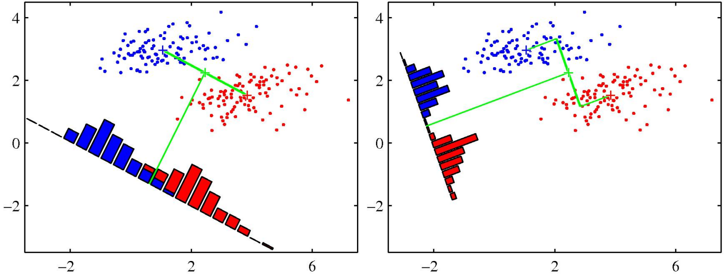
Fisher's Discriminant Analysis

Consider the linear projection:

$$y = \mathrm{w}^T \mathrm{x} \tag{1}$$

The classification rule is:

assign x to class $\mathcal{C}_1$ if $y \geq -w_0$, otherwise assign it to class $\mathcal{C}_2$.



Fisher's Discriminant Analysis

We represent the two classes using their mean vectors:

$$\mathrm{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathrm{x}_n, \qquad \mathrm{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathrm{x}_n \tag{2}$$

where N_k is the number of data points in class k, with $N = N_1 + N_2$.

A simple way to measure separation is (we use $\|\mathrm{w}\| = 1$):

$$m_2 - m_1 = \mathrm{w}^T(\mathrm{m}_2 - \mathrm{m}_1) \tag{3}$$

where $m_k = \mathrm{w}^T \mathrm{m}_k,\ k = 1, 2$ is the mean of the projected data of class $\mathcal{C}_k$.

The within-class variance of $\mathcal{C}_k$ is: $s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$ and expresses the *compactness* of class $\mathcal{C}_k$.

Fisher's Discriminant Analysis

We determine w as the one combining:

- maximization of $m_2 - m_1$ and
- minimization of $s_1^2 + s_2^2$

We obtain such a w by maximizing the *Fisher ratio*:

$$\mathcal{J}(\mathrm{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \tag{4}$$

Fisher's Discriminant Analysis

$$\mathcal{J}(\mathrm{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathrm{w}^T \mathrm{S}_B \mathrm{w}}{\mathrm{w}^T \mathrm{S}_W \mathrm{w}} \tag{5}$$

where:

$$\mathrm{S}_B = (\mathrm{m}_2 - \mathrm{m}_1)(\mathrm{m}_2 - \mathrm{m}_1)^T \tag{6}$$

$$\mathrm{S}_W = \sum_{n \in \mathcal{C}_1} (\mathrm{x}_n - \mathrm{m}_1)(\mathrm{x}_n - \mathrm{m}_1)^T + \sum_{n \in \mathcal{C}_2} (\mathrm{x}_n - \mathrm{m}_2)(\mathrm{x}_n - \mathrm{m}_2)^T \tag{7}$$

Setting $\frac{\theta \mathcal{J}(\mathrm{w})}{\theta \mathrm{w}} = 0$ we get:

$$\mathrm{w} \propto \mathrm{S}_W^{-1}(\mathrm{m}_2 - \mathrm{m}_1). \tag{8}$$

FDA and Least Squares

When we set the targets of LS-based regression to:

- $t_n = N/N_1$ for x_n belonging to class $\mathcal{C}_1$
- $t_n = -N/N_2$ for x_n belonging to class $\mathcal{C}_2$

then LS-based regression is equivalent with FDA.

The error function is:

$$E(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathrm{w}^T\mathrm{x}_n + w_0 - t_n)^2 \tag{9}$$

Using the derivatives of E w.r.t. w_0 and w and using the above t_n we get:

$$\sum_{n=1}^{N}(\mathrm{w}^T\mathrm{x}_n + w_0 - t_n) = 0 \quad \Rightarrow \quad w_0 = -\mathrm{w}^T\mathrm{m} \tag{10}$$

$$\sum_{n=1}^{N}(\mathrm{w}^T\mathrm{x}_n + w_0 - t_n)\mathrm{x}_n = 0 \quad \Rightarrow \quad \left(\mathrm{S}_W + \frac{N_1 N_2}{N}\mathrm{S}_B\right)\mathrm{w} = N(\mathrm{m}_1 - \mathrm{m}_2) \tag{11}$$

FDA and Least Squares

From the above:

$$\mathrm{w} \propto \mathrm{S}_W^{-1}(\mathrm{m}_2 - \mathrm{m}_1) \tag{12}$$

which is in the same direction as the solution of FDA.

We have also found an expression for w_0:

- x should be classified to $\mathcal{C}_1$ if $y(\mathrm{x}) = \mathrm{w}^T(\mathrm{x} - \mathrm{m}) > 0$, and to $\mathcal{C}_2$ otherwise.

Linear Discriminant Analysis

When $K > 2$ we define a projection:

$$\mathrm{y} = \mathrm{W}^T \mathrm{x} \tag{13}$$

where $\mathrm{W} \in \mathbb{R}^{D \times D'}$, with $D' > 1$.

We define the following *scatter matrices*:

$$\mathrm{S}_W = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} (\mathrm{x}_n - \mathrm{m}_k)(\mathrm{x}_n - \mathrm{m}_k)^T \tag{14}$$

$$\mathrm{S}_T = \sum_{n=1}^{N} (\mathrm{x}_n - \mathrm{m})(\mathrm{x}_n - \mathrm{m})^T \tag{15}$$

$$\mathrm{S}_B = \sum_{k=1}^{K} N_k (\mathrm{m}_k - \mathrm{m})(\mathrm{m}_k - \mathrm{m})^T \tag{16}$$

Linear Discriminant Analysis

In the above:

$$\mathrm{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathrm{x}_n \tag{17}$$

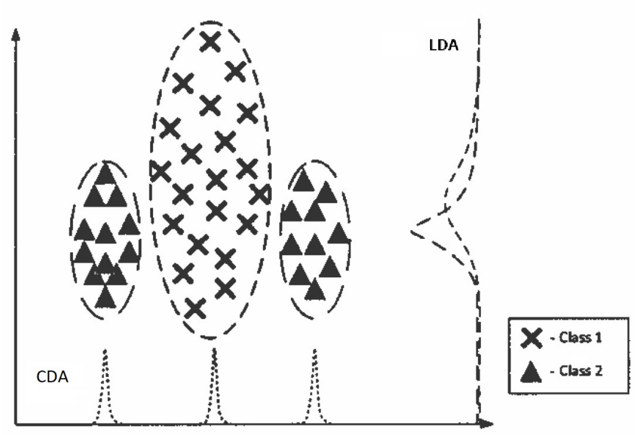
$$\mathrm{m} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{x}_j = \frac{1}{N} \sum_{k=1}^{K} N_k \mathrm{m}_k. \tag{18}$$

The projection W is determined by maximizing for:

$$\mathcal{J}(\mathrm{w}) = Tr\left\{ (\mathrm{W}^T \mathrm{S}_W \mathrm{W})^{-1} (\mathrm{W}^T \mathrm{S}_B \mathrm{W}) \right\} \tag{19}$$

and is formed by the eigenvectors corresponding to the largest $(K-1)$ eigenvalues of the matrix $\mathrm{S}_W^{-1}\mathrm{S}_B$.

Linear Discriminant Analysis: Multiple subclasses



Linear Discriminant Analysis: Multiple subclasses

Subclass Discriminant Analysis:

1. For each class $\mathcal{C}_k, \ k = 1, \dots, K$
   1. Cluster the class data $\mathcal{D}_k = \{x_n \in \mathcal{C}_k\}$ in S_k groups, each formed by N_{kj} data points
   2. Represent group j of class $\mathcal{C}_k$ by the corresponding cluster mean vector $m_{kj}, \ j = 1, \dots, N_{kj}$
2. Calculate the scatter matrices (using $p_{kj} = N_{kj}/N$):

$$\mathrm{S}_T = \sum_{n=1}^{N} (\mathrm{x}_n - \mathrm{m})(\mathrm{x}_n - \mathrm{m})^T \tag{20}$$

$$\mathrm{S}_B = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \sum_{j=1}^{N_k} \sum_{h=1}^{N_l} p_{kj} p_{lh} (\mathrm{m}_{kj} - \mathrm{m}_{lh})(\mathrm{m}_{kj} - \mathrm{m}_{lh})^T \tag{21}$$

3. The projection W is formed by the eigenvectors corresponding to the largest $(K-1)$ eigenvalues of the matrix $\mathrm{S}_T^{-1}\mathrm{S}_B$

Linear Discriminant Analysis: Multiple subclasses

Clustering-based Discriminant Analysis:

1. For each class $\mathcal{C}_k, \ k = 1, \ldots, K$
   1. Cluster the class data $\mathcal{D}_k = \{x_n \in \mathcal{C}_k\}$ in S_k groups, each formed by N_{kj} data points
   2. Represent group j of class $\mathcal{C}_k$ by the corresponding cluster mean vector $m_{kj}, \ i = 1, \ldots, N_{kj}$
2. Calculate the scatter matrices (using subclass indicator values α_n^{kj}):

$$
\begin{aligned}
\mathrm{S}_W &= \sum_{k=1}^{K} \sum_{j=1}^{N_k} \sum_{n=1}^{N} \alpha_n^{kj} (\mathrm{x}_n - \mathrm{m}_{kj})(\mathrm{x}_n - \mathrm{m}_{kj})^T \qquad (22) \\
\mathrm{S}_B &= \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \sum_{j=1}^{N_k} \sum_{h=1}^{N_l} p_{kj} p_{lh} (\mathrm{m}_{kj} - \mathrm{m}_{lh})(\mathrm{m}_{kj} - \mathrm{m}_{lh})^T \qquad (23)
\end{aligned}
$$

3. The projection W is formed by the eigenvectors corresponding to the largest $(K-1)$ eigenvalues of the matrix $\mathrm{S}_W^{-1}\mathrm{S}_B$