Prof. Dr. Rudolf Mathar, Dr. Gholamreza Alirezaei, Dr. Arash Behboodi

Exercise 8
- Proposed Solution -

Friday, January 13, 2017

Solution of Problem 1

a) (2-Classes) For the dataset with two classes, Figure 1 presents the linear discriminant rule. Note that maximum likelihood discriminant rule is the same for two classes.

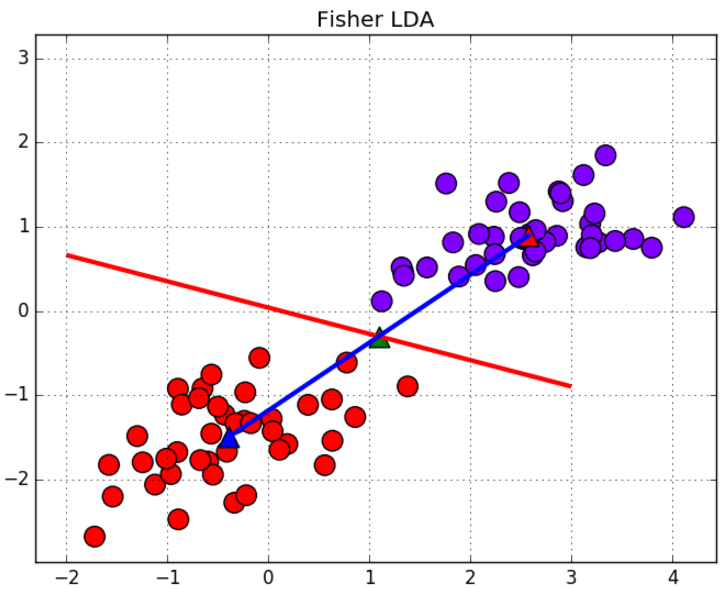

Figure 1: Dataset with two classes

The steps are as follows:

i) For each class find the class mean $\overline{\mathbf{x}}_k$; find $\overline{\mathbf{x}}$; find $\mathbf{W}$ and $\mathbf{B}$ (or $\mathbf{S}$ instead)

ii) Perform spectral decomposition of $\mathbf{W}^{-1}\mathbf{B}$ (or $\mathbf{W}^{-1}\mathbf{S}$) and find the maximum eigenvalue $\mathbf{a}$.

iii) Discriminant Rule allocates a point $\mathbf{x}$ to the group l if $|\mathbf{a}^T\mathbf{x}-\mathbf{a}^T\overline{\mathbf{x}}_l| < |\mathbf{a}^T\mathbf{x}-\mathbf{a}^T\overline{\mathbf{x}}_j|$ for all $j = 1, \dots, g$. The borders of regions are determined by $|\mathbf{a}^T\mathbf{x}-\mathbf{a}^T\overline{\mathbf{x}}_l| = |\mathbf{a}^T\mathbf{x}-\mathbf{a}^T\overline{\mathbf{x}}_j|$. These lines are given by:

$$\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\left(\frac{\overline{\mathbf{x}}_k + \overline{\mathbf{x}}_l}{2}\right) = 0.$$

b) (3-Classes) For the dataset with three classes, Figure 2 presents the linear discriminant rule and maximum likelihood discriminant rule.

The steps are as follows:

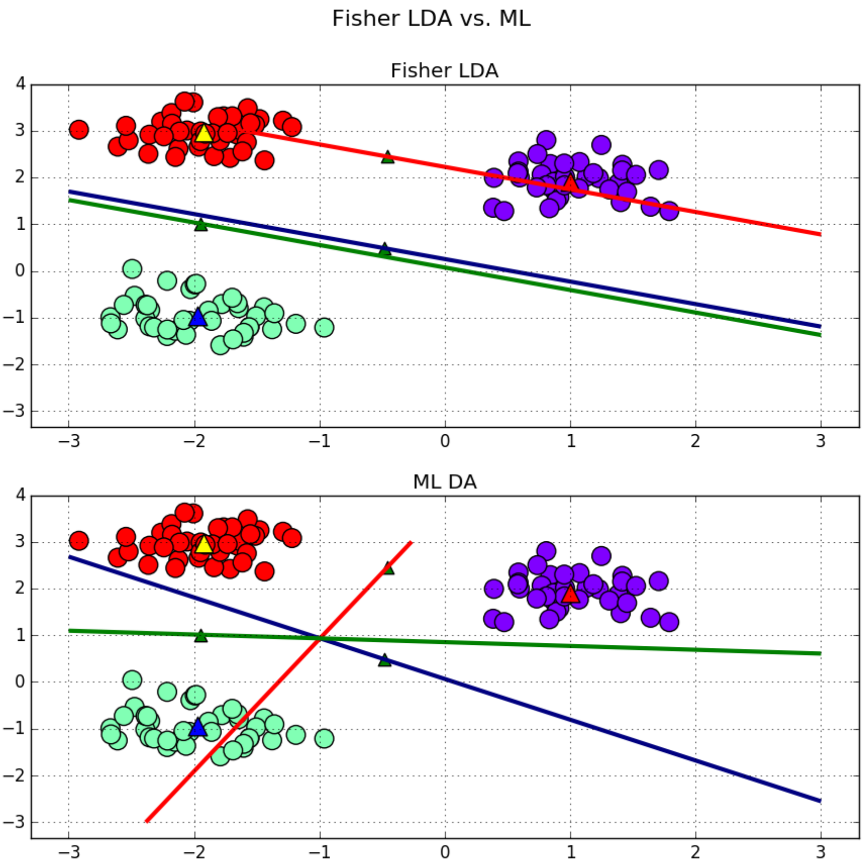

Figure 2: Dataset with three classes

i) Find the maximum likelihood estimation $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ as $\frac{\mathbf{W}}{n}$; find the class mean $\overline{\mathbf{x}}_k$.

ii) The ML rule allocates $\mathbf{x}$ to C_l which minimizes the Mahalanobis distance:

$$(\mathbf{x} - \overline{\mathbf{x}}_l)^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x} - \overline{\mathbf{x}}_l).$$

Solution of Problem 2

a) (2-Classes) For the dataset with two classes, Figure 3 presents the linear discriminant rule. Note that maximum likelihood discriminant rule is the same for two classes.

The steps are as follows:

i) For each class find the class mean $\overline{\mathbf{x}}_k$; find $\overline{\mathbf{x}}$; find $\mathbf{W}$ and $\mathbf{B}$ (or $\mathbf{S}$ instead)

ii) Perform spectral decomposition of $\mathbf{W}^{-1}\mathbf{B}$ (or $\mathbf{W}^{-1}\mathbf{S}$) and find the maximum eigenvalue $\mathbf{a}$.

iii) Discriminant Rule allocates a point $\mathbf{x}$ to the group l if $|\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\overline{\mathbf{x}}_l| < |\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\overline{\mathbf{x}}_j|$ for all $j = 1, \ldots, g$. The borders of regions are determined by $|\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\overline{\mathbf{x}}_l| = |\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\overline{\mathbf{x}}_j|$. These lines are given by:

$$\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\left(\frac{\overline{\mathbf{x}}_k + \overline{\mathbf{x}}_l}{2}\right) = 0.$$

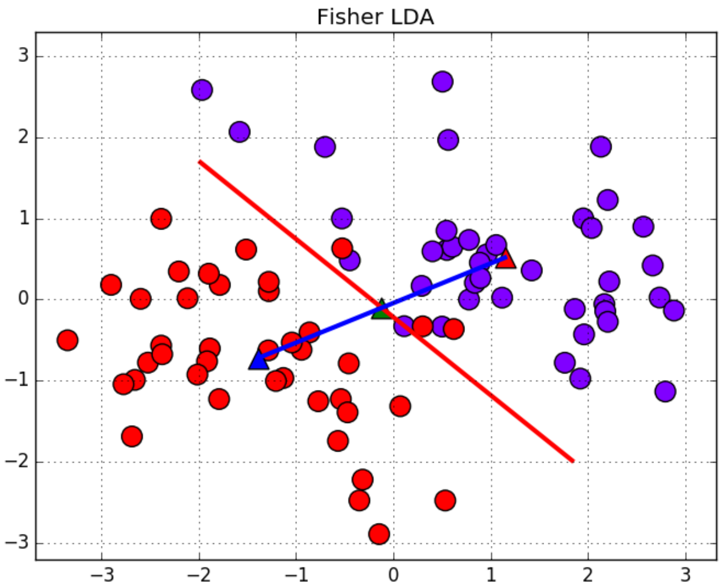

Figure 3: Dataset with two classes

b) (3-Classes) For the dataset with three classes, Figure 4 presents the linear discriminant rule and maximum likelihood discriminant rule.

The steps are as follows:

i) Find the maximum likelihood estimation $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ as $\frac{\mathbf{W}}{n}$; find the class mean $\overline{\mathbf{x}}_k$.

ii) The ML rule allocates $\mathbf{x}$ to C_l which minimizes the Mahalanobis distance:

$$(\mathbf{x} - \overline{\mathbf{x}}_l)^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x} - \overline{\mathbf{x}}_l).$$

Fisher LDA vs. ML

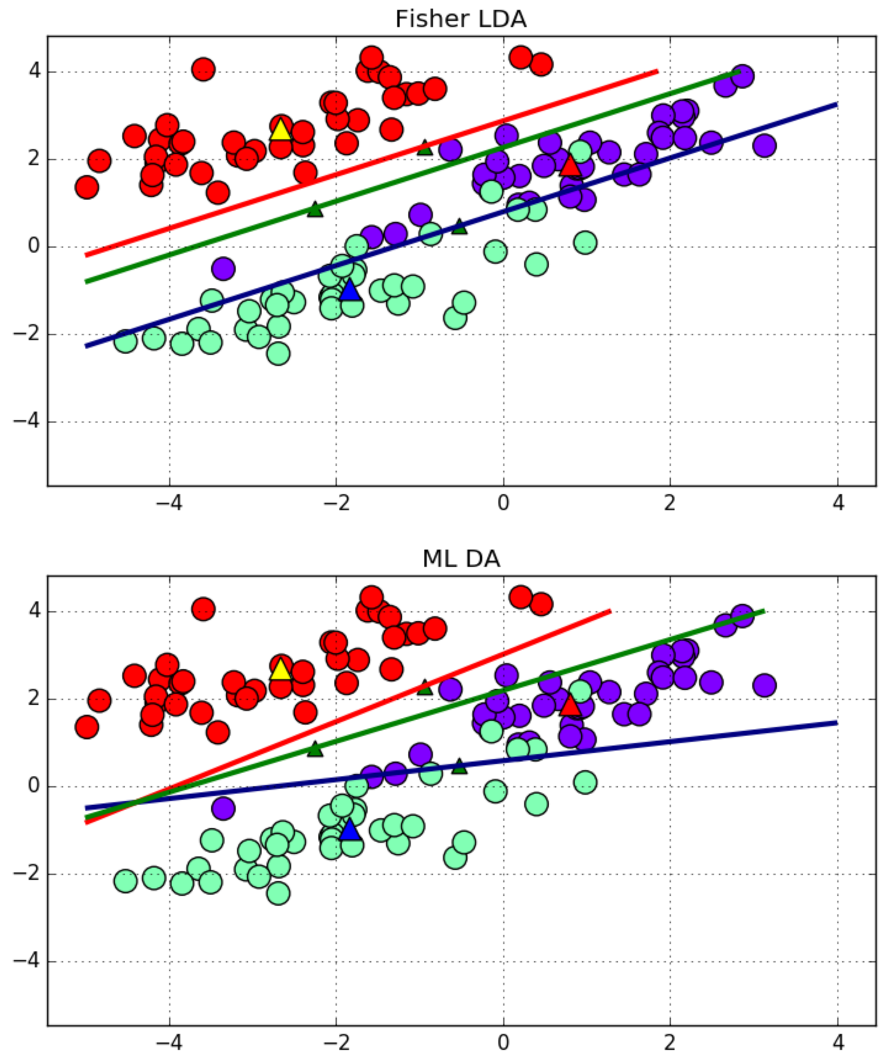

Figure 4: Dataset with three classes

Solution of Problem 3

a) If the n points are clustered into $S_1, \ldots, S_n$, then ML-cluster analysis writes as

$$\max_{S_1,\ldots,S_g} \sum_{k=1}^{g} \sum_{i \in S_k} \log f_k(\mathbf{x}_i) = \max_{S_1,\ldots,S_g} \sum_{k=1}^{g} \sum_{i \in S_k} \text{const.} - \frac{1}{2} \log |\mathbf{\Sigma}| - \frac{1}{2} \left\{ (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}.$$

Therefore having Σ and μ_k, the ML-cluster analysis is given by

$$\min_{S_1,\ldots,S_g} \sum_{k=1}^{g} \sum_{i \in S_k} \log |\mathbf{\Sigma}| + \left\{ (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}.$$

b) Given clustering of samples $S_1, \ldots, S_g$, the ML-estimation of $\mathbf{\Sigma}$ results from the minimization of above expression for fixed $S_1, \ldots, S_g$. Following similar argument from

ML estimation of covariance matrix, μ_k are estimated by $\overline{\mathbf{x}}_k$. Using these values and differentiating with respect to $\mathbf{\Sigma}^{-1}$, similar to ML-estimation of covariance matrix, the ML-estimation of $\mathbf{\Sigma}$ is given by:

$$n\hat{\mathbf{\Sigma}} = \sum_{k=1}^{g} \sum_{i \in S_k} \left\{ (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \right\} \implies \hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{W},$$

where $\mathbf{W}$ is within-group sum of squares.

c) Using the above estimation, ML-estimation can be written as

$$\min_{S_1,\ldots,S_g} \sum_{k=1}^{g} \sum_{i \in S_k} \log |\frac{\mathbf{W}}{n}| + \left\{ (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \mathbf{W}^{-1} n (\mathbf{x}_i - \overline{\mathbf{x}}_k) \right\}.$$

But:

$$\sum_{k=1}^{g} \sum_{i \in S_k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \mathbf{W}^{-1} n (\mathbf{x}_i - \overline{\mathbf{x}}_k) = \sum_{k=1}^{g} \sum_{i \in S_k} \operatorname{tr}(\mathbf{W}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T) = \operatorname{tr}(\mathbf{W}^{-1}\mathbf{W}) = p.$$

Therefore the ML-estimation can be written as:

$$\min_{S_1,\ldots,S_g} \det(\mathbf{W}).$$

d) If $\mathbf{\Sigma}$ is known, ML-cluster analysis is written as:

$$\min_{S_1,\ldots,S_g} \sum_{k=1}^{g} \sum_{i \in S_k} \log |\mathbf{\Sigma}| + \left\{ (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_k) \right\}.$$

Since Σ is known and irrelevant for the optimization, only the second term is important. Now see that from the argument used above:

$$\sum_{k=1}^{g} \sum_{i \in S_k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_k) = \operatorname{tr}(\mathbf{W}\mathbf{\Sigma}^{-1}).$$

Therefore the ML-analysis writes as:

$$\min_{S_1,\ldots,S_g} \operatorname{tr}(\mathbf{W}\mathbf{\Sigma}^{-1}).$$