Sheet 9)

Exercise 2) Classification by linear discriminant analysis. Training data set:

$T_{train} = \{(-2.1, 1), (-0.9, 1), (0.6, 2), (1.5, 2), (2.7, 2)\}$

Validation set: $\{(-1.2, 1), (0.5, 1), (1.4, 2)\}$

a) Use linear discriminant analysis to build a classifier based on the training data.

b) Evaluate the generalization error for the constructed predictor using the 0-1 loss and the validation set approach.

Solution:

a) $N_1 = 2$, $N_2 = 3$, $N = N_1 + N_2 = 5$, $r = 2$ groups

$\hat{P}_G(1) = \frac{N_1}{N} = \frac{2}{5} = 0.4$, $\hat{P}_G(2) = \frac{N_2}{N} = \frac{3}{5} = 0.6$

$\hat{\mu}_1 = \frac{1}{2}(-2.1 - 0.9) = -1.5$; $\hat{\mu}_2 = \frac{1}{3}(0.6 + 1.5 + 2.7) = 1.6$

$$\hat{\Sigma} = \frac{1}{5-2}\Big[(-2.1+1.5)^2 + (-0.9+1.5)^2 + (0.6-1.6)^2 + (1.5-1.6)^2 + (2.7-1.6)^2\Big] = 0.98$$

$$\Rightarrow \hat{\Sigma}^{-1} = \frac{50}{49} \simeq 1.02$$

Denote $h^{(g)}(x) = \ln(\hat{P}_G(g)) + x^T \hat{\Sigma}^{-1}\hat{\mu}_g - \frac{1}{2}\hat{\mu}_g^T \hat{\Sigma}^{-1}\hat{\mu}_g$

$$h^{(1)}(x) = \ln(0.4) + x \cdot \frac{50}{49} \cdot (-1.5) - \frac{1}{2} \cdot (-1.5)^2 \cdot \frac{50}{49} \simeq -1.53x - 2{,}06$$

$$h^{(2)}(x) = \ln(0.6) + x \cdot \frac{50}{49} \cdot 1.6 - \frac{1}{2} \cdot (1.6)^2 \cdot \frac{50}{49} \simeq 1.63x - 1.82$$

Classifier: $\hat{f}_b(x) \approx \underset{g \in \{1,2\}}{\operatorname{argmax}} \left(h^{(g)}(x)\right)$ where $\begin{cases} h^{(1)}(x) \simeq -1.53x - 2.06 \\ h^{(2)}(x) \simeq 1.63x - 1.82 \end{cases}$

b) $T_{val} = \{(x_i, y_i)\}_{i=1}^{3} = \{(-1.2, 1), (0.5, 1), (1.4, 2)\}$

$h^{(1)}(x_1) \simeq -0.22$, $h^{(2)}(x_1) \simeq -3.78 \Rightarrow \hat{f}_b(x_1) = 1$ (right)

$h^{(1)}(x_2) \simeq -2.83$, $h^{(2)}(x_2) \simeq -1.01 \Rightarrow \hat{f}_b(x_2) = 2$ (wrong)

$h^{(1)}(x_3) \simeq -4.20$, $h^{(2)}(x_3) \simeq 0.46 \Rightarrow \hat{f}_b(x_3) = 2$ (right)

Expected Generalization Error (EGE) with validation set:

$$EGE(f_b) = \frac{1}{|T_{val}|} \sum_{(x_i, y_i) \in T_{val}} L(y_i, f_b(x_i))$$

$$EGE(f_b) = \frac{1}{3}(0 + 1 + 0) = \frac{1}{3}$$