# ML01 – Spring 2019
# Lab 3: Linear classification

## 1 Classification of the `spam` data

1. Load the `spam` dataset. Plot the data. Which predictors seem to provide useful information for the classification ?
2. Split the data into a training set (approximately 2/3 of the data) and a test set.
3. Build a LDA classifier for this data (using function `lda` in package `MASS`). Print its confusion matrix evaluated using the test data. What is the test misclassification error rate ?
4. Using function `roc` in package pROC, plot the ROC curve of the LDA classifier built in the previous section. If we want to detect 80% of the spams, which percentage of the good emails will be incorrectly classified as spam ?
5. Build a $k$-NN classifier using the same data. Compare its performance to that of the LDA classifier, for different values of $k$.

## 2 Estimation of the Bayes error rate

We consider a classification problem with $K = 3$ classes and $p = 2$ input variables. The marginal distribution of $Y$ is defined by the following prior probabilities :

$$\pi_1 = 0.3, \quad \pi_2 = 0.3, \quad \pi_3 = 0.4,$$

and the conditional densities of $\mathbf{X}$ given $Y = k$, $k = 1, 2, 3$ are multivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with

$$\boldsymbol{\mu}_1 = (0, 0)^T, \quad \boldsymbol{\mu}_2 = (0, 2)^T, \boldsymbol{\mu}_3 = (2, 0)^T,$$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

1. Estimate de Bayes error rate for this problem (use function `dmvnorm` of package `mvtnorm` to compute the density of the multivariate normal distribution.

2. Generate training datasets of different sizes, and compare the error probability of the LDA classifier trained with this data to the Bayes error rate.