TIS02 Automne 2015
TP 1
Discriminant Analysis

1 Simulated Data

1.1 Programming

We propose to study and compare three discriminant analysis models in a binary case (datasets with $g = 2$ classes) : quadratic discriminant analysis, linear discriminant analysis, and the naive Bayes classifier.

1. Recall the expression for the parameter estimates in each case.
2. Fill in the blanks in the functions adq.app, adl.app, nba.app and ad.val in the anadisc.R file.

The functions adq.app, adl.app and nba.app allow to train the models : they accept as inputs the data array Xapp and the vector zapp of associated labels, and return the estimated model parameters (proportions, vectors of averages and covariance matrices of the classes).
They make use of the array structure, which makes it possible to create arrays with more than two dimensions, in order to store the covariance matrices.

The function ad.val computes the posterior probabilities for a given dataset, and then classifies the data according to these probabilities : it accepts thus the model parameters and the dataset Xtst to classify, and returns a structure containing the estimated posterior probabilities prob and the associated classes.
This function uses the function mvdnorm, which makes it possible to compute the density of a multivariate Gaussian distribution for a given dataset.
It will be possible to use the lcur.ad function, in order to plot the level curves of the estimated posterior probabilities $\widehat{\mathbb{P}}(\omega_1|\boldsymbol{x})$. The decision boundary corresponds to the level curve $\widehat{\mathbb{P}}(\omega_1|\boldsymbol{x}) = 0.5$.

1.2 Testing the functions

We want to compare the performances of the discriminant analysis models on simulated data. Interpret the script DataGen.R : what is its purpose ? We will use the script to generate data with controlled characteristics (class parameters).
For a given dataset, we will observe the following procedure :

1. separate the data into a training set and a test set ;
2. learn the class parameters on the training set,
3. classify the test data and compute the associated error rate.

We can repeat this procedure 100 times in order to compute the average error rate and confidence intervals.

2 Real data

2.1 "Pima" data

We want to apply the three discriminant analysis models to the prediction of diabetes in a population of Indians. We can load the data using the following code :

```
Donn <- read.csv("Pima.csv", header=T)
X <- Donn[,1:7]
z <- Donn[,8]
```

We can then use the same protocole as for the simulated data — do not forget to repeat the procedure (training data selection, test data classification) 100 times. Compute the average error rates. What can you observe ? Why ?

2.2 "Breast cancer Wisconsin" data

We now consider a classification problem consisting in predicting the level of gravity of a tumor according to some of its descriptive features. We can load the data using the following code :

```
Donn <- read.csv("bcw.csv", header=T)
X <- Donn[,1:9]
z <- Donn[,10]
```

Repeat the procedure (training data selection, test data classification) $N = 100$ times. Compute the average error rates. What can you observe ? Why ?