# Lab 2 : more on supervised learning

January 16, 2017

## 1 Linear discriminant analysis (LDA)

**Theoritical aspect**
Let us consider two Gaussian populations in $\mathbb{R}^p$ with the same covariance structure. We have observations drawn from a mixture of these two populations. The conditional distributions of $X$ given $Y = +1$ (respectively $Y = -1$) are multivariate Gaussian distributions $\mathcal{N}_p(\mu_+, \Sigma)$ (respectively $\mathcal{N}_p(\mu_-, \Sigma)$). We denote their respective probability density functions $f_+$ and $f_-$. The two vectors $\mu_+$ and $\mu-$ both belong to $\mathbb{R}^p$ and $\Sigma$ is a symmetric matrix. We also denote $\pi_+ = \mathbb{P}[Y = +1]$. We recall that the p.d.f. of $\mathcal{N}_p(\mu; \Sigma)$ reads :

$$f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

and that the covariance matrix of a random vector $X$ is defined as

$$\Sigma = \mathbb{E}\left((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T\right)$$

1. Use the Bayes formula to compute $\mathbb{P}[Y = +1|X = x]$, $\mathbb{P}[Y = -1|X = x]$, as function of $f_+$, $f_-$ and $\pi_+$.

2. Express the log-ratio of the two classes :

$$log\left(\frac{\mathbb{P}[Y = +1|X = x]}{\mathbb{P}[Y = -1|X = x]}\right)$$

   in function of $\mu_+$, $\mu_-$, $\pi_+$ and $\Sigma$.

3. We have some observations drawn from this mixture and we assume that $\mu_+$, $\mu_-$, $\pi_+$ and $\Sigma$ are unknown. We assume that the sample contains $n$ observations $\{(x_1, y_1), \cdots, (x_n, y_n)$ and that $\sum_{i=1}^n 1_{\{y_i=+1\}} = m$. Use the moments method to propose parametric estimators of the unknown parameters.

4. Justify the following choice of the classifier

$$\begin{cases} 1 \text{ if } x^T\widehat{\Sigma}^{-1}(\widehat{mu}_+ - \widehat{\mu}_-) > \frac{1}{2}\widehat{mu}_+\widehat{\Sigma}^{-1}\widehat{mu}_+ - \frac{1}{2}\widehat{mu}_-\widehat{\Sigma}^{-1}\widehat{mu}_- + log(1 - m/n) - log(m/n) \\ -1 \text{ otherwise} \end{cases}$$

5. What happens when the two covariance matrices differ

6. How can we generalize the linear discriminant analysis to the multiclass setting?

**LDA in practice**

We now apply LDA on synthetic data and thereafter to real data. In this last case, we split randomly the dataset into two parts : a training set (around 70% of the data) and a validation set (the 30% remaining).

1. Import `sklearn` package
   `from sklearn.lda import LDA`

2. Create a LDA model
   `my_lda = LDA()`

3. Learn the model from the data `dataX` and their corresponding labels `dataY`
   my_lda.fit(dataX,dataY)

4. Apply the LDA on the mixture generated by the function `rand_bi_gauss`. Estimate the prediction error using the test sample