Master M2 MVA 2019/2020 - Graphical models - Homework 1

These exercises are due on or before November 22th 2019 and should be submitted on Moodle. They can be done in groups of two students. The write-up can be in English or in French. Please submit your answers as a pdf file that you will name MVA DM1 <your name>.pdf if you worked alone or MVA DM1 <name1> <name2>.pdf with both of your names if you work as a group of two. Indicate your name(s) as well in the documents. Please submit your code as a separate zipped folder and name it MVA DM1 <your name>.zip if you worked alone or MVA DM1 <name1> <name2>.zip with both of your names if you worked as a group of two. Note that your files should weight no more than 16Mb.

1- Learning in discrete graphical models

Consider the following model: z and x are discrete variables taking respectively M and K different values with $p(z = m) = \pi_m$ and $p(x = k|z = m) = \theta_{mk}$.

Compute the maximum likelihood estimator for π and θ based on an i.i.d. sample of observations. Please provide your derivations and not just the final answer.

2- Linear classification

The files trainA, trainB and trainC contain samples of data (x_n, y_n) where $x_n \in \mathbb{R}^2$ and $y_n \in \{0, 1\}$ (each line of each file contains the 2 components of x_n then y_n). The goal of this exercise is to implement linear classification methods and to test them on the three data sets. The code should be written in R or Python. The source code should be handed in along with results. However, all the requested figures should be printed on paper or part of a pdf file which is turned in, with clear titles that indicate what the figures represent. Therefore, we recommend to write your code and report thanks to a Markdown file (in R or Python). The discussions may of course be handwritten.

1. Generative model (LDA)

Given the class variable, the data are assumed to be Gaussian with different means for different classes but with the same covariance matrix.

$$y \sim \text{Bernoulli}(\pi), x|y = i \sim \text{Normal}(\mu_i, \Sigma).$$

a. Derive the form of the maximum likelihood estimator for this model.

b. What is the form of the conditional distribution p(y = 1|x)? Compare with the form of logistic regression.

c. Implement the MLE for this model and apply it to the data. Represent graphically the data as a point cloud in $\mathbb{R}^2$ and the line defined by the equation

$$p(y = 1|x) = 0.5.$$

2. Logistic regression

Implement logistic regression for an affine function $f(x) = w^\intercal x + b$ (do not forget the constant term).

a. Give the numerical values of the parameters learnt.

b. Represent graphically the data as a cloud point in $\mathbb{R}^2$ as well as the line defined by the equation

$$p(y=1|x)=0.5.$$

3. Linear regression

Consider class y as a real valued variable taking the values 0 and 1 only. Implement linear regression (for an affine function $f(x)=w^\intercal x+b)$ by solving the normal equations.

a. Provide the numerical values of the learnt parameters.
b. Represent graphically the data as a point cloud in $\mathbb{R}^2$ as well as the line defined by the equation

$$p(y=1|x)=0.5.$$

4. Application

Data in the files testA, testB and testC are respectively drawn from the same distribution as the data in the files trainA, trainB and trainC. Test the different models learnt from the corresponding training data on these test data.

a. Compute for each model the misclassification error (i.e. the fraction of the data misclassified) on the training data and compute it as well on the test data.
b. Compare the performances of the different methods on the three datasets. Is the misclassification error larger, smaller, or similar on the training and test data? Why? Which methods yield very similar/dissimilar results? Which methods yield the best results on the different datasets ? Provide an interpretation.

5. QDA model

We finally relax the assumption that the covariance matrices for the two classes are the same. So, given the class label the data are assumed to be Gaussian with means and covariance matrices which are a priori different.

$$y\sim\text{Bernoulli}(\pi),x|y=i\sim\text{Normal}(\mu_i,\Sigma_i).$$

Implement the maximum likelihood estimator and apply it to the data. (a) Provide the numerical values of the learnt parameters.

b. Represent graphically the data as well as the conic defined by

$$p(y=1|x)=0.5.$$

c. Compute the misclassification error for QDA for both train and test data.
d. Comment the results as previously.