Master M2 MVA 2018/2019 - Graphical models - HWK 1

- These exercises are due on or before October 24th 2017.
- They should be submitted on the Moodle.
- **They can be done in groups of two students**.
- The write-up can be in English or in French, can be typed or scanned (except for the figures).
- **Please follow precisely the formatting described in Section 3.**
- 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 succinctly your derivations and not just the final answer.

2 Linear classification

The files `classificationA.train`, `classificationB.train` and `classificationC.train` 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 choice of the programming language is yours (we however recommend Python, Matlab, Scilab, Octave, or R). The source code should be handed in along with results.

Do not forget to use the page formatting imposed in Section 3.

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.

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

   (a) Derive the form of the maximum likelihood estimator for this model. *Indication* : the model was presented in class but not the MLE computations. Compare $p(y = 1|x)$ with the form of logistic regression.

(b) 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 $f(x) = w^\top x + b$ (do not forget the constant term), using the IRLS algorithm (Newton-Raphson) which was described in class.

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

   (a) 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. Data in the files `classificationA.test`, `classificationB.test` and `classificationC.test` are respectively drawn from the same distribution as the data in the files `classificationA.train`, `classificationB.train` et `classificationC.train`. 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.

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

   Implement the maximum likelihood estimator and apply it to the data.

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

   (b) Represent graphically the data as well as the conic defined by

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

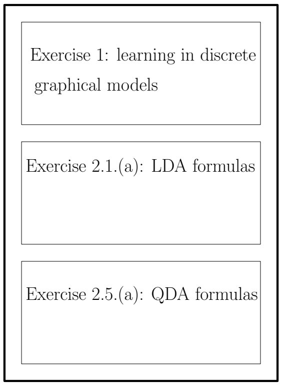
   (c) Compute the misclassification error for QDA for both train and test data.

   (d) Comment the results as previously.

3 Formatting

In order to reduce the correction time of these homeworks, please follow the following formatting precisely :

- Page 1

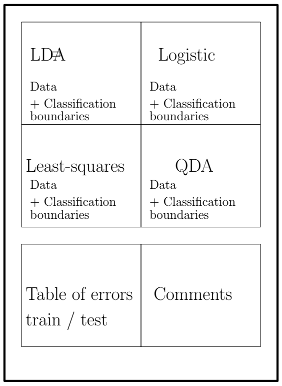


Exercise 1: learning in discrete graphical models

Exercise 2.1.(a): LDA formulas

Exercise 2.5.(a): QDA formulas

- Page 2 : dataset A



LDA
Data
+ Classification boundaries

Logistic
Data
+ Classification boundaries

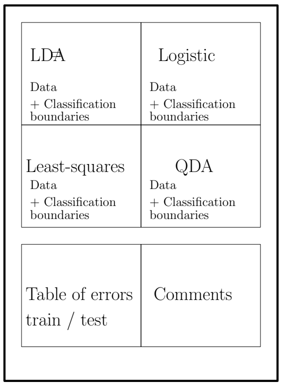
Least-squares
Data
+ Classification boundaries

QDA
Data
+ Classification boundaries

Table of errors train / test

Comments

- Page 3 : dataset B



LDA
Data
+ Classification boundaries

Logistic
Data
+ Classification boundaries

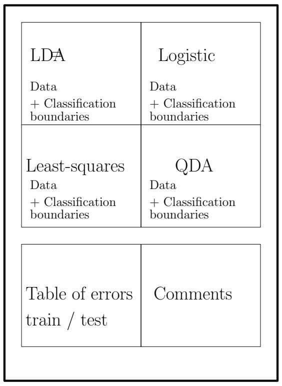
Least-squares
Data
+ Classification boundaries

QDA
Data
+ Classification boundaries

Table of errors train / test

Comments

- Page 4 : dataset C


LDA
Data
+ Classification boundaries
Logistic
Data
+ Classification boundaries
Least-squares
Data
+ Classification boundaries
QDA
Data
+ Classification boundaries
Table of errors train / test
Comments

- Page 5 : add detailed proofs for Page 1 if they do not fit.