Introduction to Machine Learning

Home Assignment 2

This homework is due by **Friday May 20, 2020**. It is to be returned by email to raphael.berthier@inria.fr as a **pdf** report of **maximum 3 pages** together with the ipython notebook used for the code. The results and the figures must be included into the pdf report but not the code.

This project is the continuation of the first homework **with a smaller training set** which can be downloaded at `http://www.di.ens.fr/appstat/spring-2020/project/data2.zip`. The zip archive contains two folders:

- train: contains $n = 900$ labeled images of three classes "A", "B" and "C" (300 each)
- test: contains $n_1 = 750$ labeled images (250 for each of the three classes).

The goal is to classify if an image X_i corresponds to the letter "A": i.e., the output is $Y_i = 1$ if image i is "A" and -1 otherwise (if the image is "B" or "C").

1. **Regularization of logistic regression.** Using the gradient descent algorithm implemented in homework 1 solve the following minimization problem (up to small enough precision):

$$\widehat{\beta}_\lambda \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-Y_i \langle \beta, X_i \rangle}\right) + \lambda \|\beta\|_1 \,. \tag{1}$$

   (a) Why is it often important to regularize?

   (b) Plot the test and training classification errors with the 0-1 loss associated with $\widehat{\beta}_\lambda$ as a function of λ.

   (c) What would be the best value for λ? How would you tune it?

   (d) Plot as images the estimators $\widehat{\beta}_\lambda$ for four values of λ.

   (e) Repeat questions 1.a-c) by replacing the ℓ_1 regularization with a ℓ_2 regularization $\lambda\|\beta\|_2^2$.

2. **Linear discriminant analysis (LDA).** We consider another classification model called LDA that models the data of each class as a Gaussian distribution: $\mathcal{N}(\mu_1, \Sigma)$ for the class 1 (i.e., $Y = 1$) and $\mathcal{N}(\mu_{-1}, \Sigma)$ for class -1 (i.e., $Y = -1$). The covariance matrix Σ is assumed to be the same for both classes [1]. We denote by $\pi = \mathbb{P}(Y = 1)$ the probability of a sample to get label 1.

   (a) Show that the conditional model of $\mathbb{P}(Y = 1|X)$ associated with this generative model is of the form

$$\mathbb{P}(Y = 1|X) = \frac{1}{1 + \exp\big(-\beta_0 - \langle \beta, X \rangle\big)} \,,$$

   for some $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$ depending on Σ, μ_{-1}, μ_1, and π to be explicited.

   (b) Show that, assuming $\beta_0 = 0$ (i.e., no intercept), the maximum likelihood of β of this probabilistic model corresponds to the solution of logistic regression (1) with $\lambda = 0$.

   (c) Apply this model to the data by using the maximum likelihood principle to fit the parameters μ_1, μ_{-1}, Σ, and π. If the results are unstable, try to add $\lambda_0 I$ with a small λ_0 to the estimated covariance matrix. Can it be considered as a regularization?

[1] The method that allows different covariance matrices is called Quadratic Discriminant Analysis (QDA)

(d) Give the definitions of false positive, false negative and give the confusion matrix associated with LDA. What is the advantage of the confusion matrix over the classification error?

3. **Unsupervised learning.** Here, we assume that we do not have access to a training set with labeled images. We have only access to the unlabeled test images. The goal is to build an algorithm that can automatically group the images into three classes that hopefully correspond to A, B and C.

   (a) Implement from scratch the K-means algorithm with three clusters and 10 random initializations. Implement a PCA (Principal Component Analysis) of the dataset and plot the results of the K-means algorithm by projecting the images and the centers of the classes on the two principal components.

   (b) Try to associate each of the three clusters with a class A, B and C and give the corresponding confusion matrix. Comment.