Machine Learning 2020 – Homework

For parts II and III, the report should include all the codes to generate the results and the figures. The code can be in Matlab, Python or Julia and should be reasonably commented. If using Python or Julia, the preferred output is a Jupyter notebook. You can use machine learning libraries (for instance scikit-learn if using Python) or write your own functions. A pdf version of the report should also be provided to prevent any difficulty with reading the files.

The report is due for April 30 and should be sent to `olivier.rivoire@college-de-france.fr` with subject [ML homework]. A penalty will be applied for each day of delay. A maximum of two students can send a common report.

I. COURSE QUESTIONS

A. Maximum likelihood estimation

Alice throws a coin 100 times and obtains 55 times a tail. Estimate by maximum likelihood the probability that the coin gives a tail. What confidence do we have in this result? Should Alice consider the coin to be unfair?

B. Bayesian estimation

Bob is tested for a disease. The test, which is either positive or negative, is only 90% reliable. Given that 1% of people of Bob's age and background have the disease, what is the probability that Bob has the disease?

By redesigning the test, you can either reduce from 10% to 5% the false positive rate (less negative results when the patient is positive) or reduce from 10% to 5% the false negative rate (less positive results when the patient is negative): what is preferable?

C. Information theory

The binary erasure channel is a discrete memoryless channel where each input $x_i \in \{0, 1\}$ is either transmitted reliably, with probability $1-\epsilon$, or replaced by an error symbol $*$, with probability ϵ. What is the capacity of this channel?

More generally, a memoryless erasure channel takes inputs from an alphabet of q symbols $\{0, 1, 2, \dots, q\}$: any of these symbols is transmitted reliably with probability $1-\epsilon$ and replaced by an error symbol $*$ with probability ϵ. What is the capacity of this channel?

D. Maximum entropy method

Consider N binary sequences of length p: $\sigma_{ij} = \pm 1$ with $i = 1, \dots, N$ and $j = 1, \dots, p$. We use the maximum entropy method to estimate $P(\sigma_1, \dots, \sigma_p)$. Show that if we choose to constrain for each j the average of σ_j to the empirical mean $\mu_j = \sum_i \sigma_{ij}/N$, the maximum entropy principle leads to a distribution of the form

$$P(\sigma_1, \dots, \sigma_p) = \frac{e^{\sum_j h_j \sigma_j}}{Z}. \qquad (1)$$

What are the values of h_j and Z? What if we take $\sigma_{ij} \in \{0, 1\}$ instead of $\sigma_{ij} \in \{-1, 1\}$?

With $\sigma_j \in \{0, 1\}$, μ_j is the empirical frequency f_j at which $\sigma_j = 1$ occurs. What if we also constrain the pair frequencies f_{jk} at which $\sigma_j = 1$ and $\sigma_k = 1$ co-occur for every pair (j, k)? What if we additionally constrain the three-way frequencies $f_{jk\ell}$ for every (j, k, ℓ)? And if we constrain all n-way frequencies for $n = 1, \dots, p$?

How does Eq. (1) generalize to sequences where the variables take more than 2 values: $\sigma_j \in \{1, 2 \dots, q\}$ with $q > 2$?

II. LASSO REGRESSION AND MODEL SELECTION: WHAT MAKES A GOOD WINE?

For this exercise, data is to be downloaded from `http://www3.dsi.uminho.pt/pcortez/wine/`. It includes:
• two datasets in comma-separated value (csv) format, winequality-red.csv and winequality-white.csv
• a text file describing the content of the datasets, winequality-names.txt
• an article with more details and an analysis of the data with different methods

A. Pre-processing

Download and read the datafile winequality-red.csv. Display as an histogram the distribution of values of each of the 12 features. Notice that some features (e.g. sulphates) have outliers. They may correspond to measurement errors, which have the potential to distord the results of the analysis. We decide to exclude these outliers by retaining only the samples where each feature is within 3 standard deviations of its mean value. Clean the data with this criterion to obtain a dataset with 1451 samples. We will use this cleaned data for all the subsequent analyses.

B. Principal component analysis

Consider all features together ($p = 12$) and perform a principal component analysis on the data after normalizing each feature to zero mean and unit variance. Display the features in two dimensions by projecting them along the top two components. What feature(s) appear most correlated/anti-correlated to quality in this representation?

C. Multivariate linear regression

Next we consider all input features together and perform multivariate linear regression of the quality score against the $p = 11$ input features x. Compute $\hat{\beta}$ and R^2 after subtracting the mean for each feature so that $\bar{x}_j = 0$. Repeat after also normalizing to unit variance, so that $\bar{x}_j = 0$ and $\overline{x_j^2} = 1$. Compare the two results and comment.

D. Lasso regression and model selection

Here we want to select a few features as good predictors of quality. We divide the initial dataset in two subsets: a training set of size $N = 500$ and a test set of size 951 containing the rest of the samples (after cleaning).

Implement lasso regression, that is, solve on the training set the optimization problem

$$\min_{\beta} \left[\frac{1}{N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] . \tag{2}$$

for different values of λ. This can be done by implementing the cyclical coordinate descent algorithm presented in Lecture 1. Display the different $\hat{\beta}_j$ as a function of λ. The graph should be similar to the one we saw for the crime data (Slides 1), except that we are plotting the $\hat{\beta}_j$ as a function of λ rather than $\|\hat{\beta}(\lambda)\|_1 / \|\hat{\beta}(0)\|_1$. Comment the results.

To find an optimal value for the hyperparameter λ, implement K-fold cross-validation with $K = 10$. For different values of λ, compare the training, validation and test errors. Conclude.

If you had to measure just a few physicochemical properties to estimate quality, which would you pick? What is the test error given your choice?

E. Classification

Can we recognize red from white wines from their physicochemical properties? Using the two datasets winequality-red.csv and winequality-white.csv, propose and implement a method that classifies a wine as white or red based on its physicochemical properties (excluding quality). How accurate is your classifier?

III. CLUSTERING OF HANDWRITTEN DIGITS

Here we consider the Handwritten Digits Data Set described at `https://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits`. More specifically, we consider the datafile named optdigits.tes available at `https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/`. The same folder contains a file with a description of the format of the data.

The goal is to cluster the data without taking into account the labels (unsupervised learning). The last attribute at the end of every line in the data file, which indicates the label, should therefore be ignored. We shall use the labels only after the clustering is performed, to assess the results.

A. K-means clustering

Read the data (wihout labels) and process it to rescale the attributes to zero mean and unit variance. Take $K = 10$ and cluster the data by the K-means algorithm. Show that the result depends on the initial condition. Try 10 initial conditions and retain the best result.

Now consider the labels to assess the accuracy of the clustering. As a score, take the fraction of pairs of samples that are correctly partitioned: two samples are considered correctly partitioned if they are in the same cluster and have same the label or if they are in different clusters and have different labels. What is this score for the best clustering that you found? When you compare the results obtained from 10 different initial conditions, do you verify that your best clustering is also the one with best score based on the labels?

Try different values of K: can we infer that the digits fall into $K = 10$ categories if we do not have this information before-hand?

B. Hierarchical clustering

Take at random 25 samples labeled as 0, 25 labeled as 1 and 25 labeled as 2. Cluster these samples by agglomerative hierarchical clustering using different linkage methods (single linkage, complete linkage,. . .) and visualize the results by means of dendrograms. Are the results consistent with the labels?