MA2823: Introduction to Machine Learning
CentraleSupélec

Assignment 1

Instructor: Fragkiskos Malliaros

Due: **October 21, 2018 at 23:00**

How to submit: Please complete the first assignment **individually**. *Typeset* all your answers (**PDF** file only). Submissions should be made on **gradescope** (Assignment 1; Entry Code: MNXK6R) – use your full name (same as the one you have used for the project proposal). Make sure that the answer to each question is on a separate page (questions 1-9). For Question 9, please include in your report the basic parts of the Python code. No late assignments will be accepted.

I. The Learning Problem

Question 1 [4 points]

Which of the following problems are best suited for Machine Learning? Briefly justify your answer.

(a) [1 p] Classifying numbers into primes and non-primes.

(b) [1 p] Detecting potential fraud in credit card charges

(c) [1 p] Determining the time it would take a falling object to hit the ground.

(d) [1 p] Determining the optimal cycle for traffic lights in a busy intersection.

Question 2 [6 points]

A scientist writes a computer program that automatically determines whether a newspaper article is about science policy based on the number of times the article contains the words "science", "public", "open", "university", "government", "funding", "education", "justice", "law". What kind of machine learning problem is the above one? Explain briefly the machine learning pipeline that you will be following to deal with this problem. *[Keep your answer short; max. 10 lines]*

Question 3 [6 points]

Are the following machine learning problems? If yes, of what type? What are the design (data) matrix and, if appropriate, the target vector? *[Keep your answer short]*

(a) [3 p] Given a car owner's manual and the price of gas at the nearest gas station, predict how much it will cost to fill up the cars tank.

(b) [3 p] Compute a house's heating load (i.e., the amount of energy that is needed to maintain the temperature) from its building plan and a civil engineer's records of plans and heating loads for houses in the same neighborhood.

II. Dimensionality Reduction

Question 4 [20 points]

Let $\mathbf{M}_{m \times n}$ be a data matrix (m observations (i.e., data points), n dimensions (i.e., features)).

(a) [2 p] Are the matrices $\mathbf{M}\mathbf{M}^\top$ and $\mathbf{M}^\top\mathbf{M}$ symmetric, square and real? Justify your answer.

(b) [6 p] Show that the eigenvalues of $\mathbf{M}\mathbf{M}^\top$ are the same as the ones of $\mathbf{M}^\top\mathbf{M}$. Are their eigenvectors the same too? Justify your answer.

(c) [6 p] SVD decomposes the matrix $\mathbf{M}$ into the product $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\mathbf{\Sigma}$ is a diagonal matrix. Given that $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, write a simplified expression of $\mathbf{M}^\top\mathbf{M}$ in terms of $\mathbf{V}$, $\mathbf{V}^\top$ and $\mathbf{\Sigma}$. Can we find an analogous expression for $\mathbf{M}\mathbf{M}^\top$?

(d) [6 p] What is the relationship (if any) between the eigenvalues of $\mathbf{M}^\top\mathbf{M}$ and the singular values of $\mathbf{M}$? Justify your answer.

Question 5 [20 points]

Consider 3 data points in the 2-d space: $(-1, -1)$, $(0, 0)$, and $(1, 1)$.

(a) [9 p] We perform PCA on the data points. What is the first principal axis (write down the actual vector)?

(b) [8 p] If we project the data points into the 1-d subspace defined by the first principal axis, what are the coordinates of the data points in the 1-d subspace? In other words, find the first principal component of the data.

(c) [3 p] What is the variance of the projected data?

III. Model Evaluation and Selection

Question 6 [10 points]

We evaluated two algorithms on a task consisting in classifying mushrooms between poisonous and edible based on some descriptors. We obtained the two following confusion matrices:

	Labeled Edible	Labeled Poisonous
Edible	100	0
Poisonous	3	97

Table 1: Confusion table of Algorithm 1.

	Labeled Edible	Labeled Poisonous
Edible	96	4
Poisonous	0	100

Table 2: Confusion table of Algorithm 2.

(a) [7 p] Compute the accuracies of Algorithm 1 and Algorithm 2.

(b) [3 p] For the task of identifying poisonous mushrooms, which algorithm is better? Explain your answer.

Question 7 [4 points]

You have to implement a fraud detection system for a bank. Undetected frauds are quite costly to the bank, compared to establishing that a transaction was, in fact, not fraudulent. Which one of the following do you want to minimize? Briefly justify your answer.

(a) False positive rate

(b) False negative rate

(c) True positive rate

IV. Linear Regression, Logistic Regression and Feature Selection

Question 8 [15 points]

Let $\{y_i, X_i\}_{i=1}^m$ denotes a set of m observations, where each X_i is an n-dimensional vector. In *Ridge Regression*, a regularization term is added in the linear regression model in order to penalize the model complexity, leading to the following optimization problem:

$$\arg\min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\| + \lambda\|\theta\|_2^2,$$

where $\lambda > 0$ is a regularization parameter.

(a) [12 p] Find the closed form solution of the ridge regression problem.

(b) [3 p] Explain briefly why the ridge regression estimator is more robust to overfitting compared to the least-squares regression.

Question 9 [15 points]

The goal of this question is to examine the effect of feature selection techniques in classification. In particular, we will use the built-in implementation of `scikit-learn` for both feature selection techniques, and for the logistic regression classifier.

Description of the dataset. The dataset (`data.csv` (training set) and `test.csv` (test set)) describes a set of 102 molecules, of which 39 are judged by human experts to be musks (class 1) and the remaining 63 molecules are non-musks (class 2). The goal is to learn to predict whether new molecules will be musks or non-musks. However, the 166 features that describe these molecules depend upon the exact shape, or conformation, of the molecule. Because bonds can rotate, a single molecule can adopt many different shapes. To generate this data set, all the conformations of the molecules were generated to produce 6598 conformations. Then, a feature vector was extracted that describes each conformation. That way, the final dataset has 6598 instances and 166 features.

How to load the data. In the first part of the pipeline, you will need to load the data and to extract the class labels. That way, variable X will contain the actual data and variable Y the class labels.

```
# Load the data set
data = loadtxt('data.csv', delimiter=',')
#load 1st column
Y = data[:,0:1]
```

```
# load columns 2 - end
X = data[:,1:data.shape[1]]
```

Tasks to be done.

1. Initially, train and test the logistic regression classifier[1] without feature selection (check the various examples within the manual of `sklearn.linear_model.LogisticRegression` for how to use the classifier). Plot the *ROC* curve[2] and examine the area under curve[3] (on the test set), as well as the running time required for training (for the running time, you can use the `timeit` module[4]). Report the results. (Note that, in the training phase you should use the training part of the dataset (`data.csv`), while in the test phase, you should use the test set (`test.csv`)).

2. Apply feature selection using the *recursive feature elimination* (RFE) method[5] to keep the "most important" features, with respect to the logistic regression classifier. In this particular case, the `estimator` used by the RFE method will be the logistic regression model. For different number of features in the range of $\{20, 40, 60, 80, 100, 150\}$, examine the performance of the classifier on the test set by plotting the area under curve vs. the number of features, as well as the running time (for training) vs. the number of features. Does feature selection improve the accuracy? What do you observe in the running time? (Note that, in the test phase, you should retain only those features from the test set that are indicated by the feature selection task).

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[2] `http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html`

[3] `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html#sklearn.metrics.roc_auc_score`

[4] `https://docs.python.org/2/library/timeit.html`

[5] `http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html`