Lab 8. Classification experiment

1. Introduction

1.1. Goal

In this Lab you are going to make an experiment of the classification of a dataset using different classifiers. In this case, we provide you a dataset and your goal will be to try with different classifiers to find the one that yields the best possible results.

This lab is not guided at all, but you will have to decide which classifiers you assess and with which parameters each one. For example, focusing on the classifiers with which we have worked in the labs:

- In kNN you can modify the value of the parameter 'k', and the distance metric that is used.
- In Logistic Regression, scikit-learn allows you to change the solver or the regularization strength C.
- In SVM, you can assess with different kernels (some of them with their own parameters, such as gamma in the case of the rbf kernel), or the parameter C.

Anyway, **we hope you research about how to use other classifiers** to assess them, even though you have not used them in the lab (such as Naïve Bayes, Multilayer Perceptrons, Decision Tree, Random Forest, …) 😊.

At the end of the lab you have to deliver:

1. A **report** explaining what classifiers you have assessed, with the parameters you have assessed, how the dataset was split, the results you have obtained, etc.
2. All the code you have used to make the experiments.

In this lab we are not providing any initial code, but you can use the scikit-learn versions of previous labs **as a base** to start working.

1.1. Problem to solve

The dataset we are going to use is called the Pima Indians Diabetes dataset. It contains medical diagnostic data for predicting diabetes among female of at least 21 years old of Pima Indian heritage. The dataset was originally presented in the paper

> Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

which can be found at the link: https://pmc.ncbi.nlm.nih.gov/articles/PMC2245318/

The dataset is provided in a csv file, called Pima_indian_diabetes.csv. It consists of 768 observations (rows) with 8 attributes, each one in a different column. The attributes are:

1. **Pregnancies**: Number of times pregnant.
2. **Glucose**: Plasma glucose concentration after 2 hours in an oral glucose tolerance test.
3. **BloodPressure**: Diastolic blood pressure (mm Hg).
4. **SkinThickness**: Triceps skinfold thickness (mm).
5. **Insulin**: 2-hour serum insulin level (mu U/ml).
6. **BMI**: Body Mass Index (weight in kg/(height in m)^2).
7. **DiabetesPedigreeFunction**: Diabetes pedigree function, indicating genetic predisposition.
8. **Age**: Age in years.

Moreover, in the csv provided with this assignment, there is a 9th column called **Outcome**, which stands for the diagnosis of each patient (0 indicates non-diabetic and1 indicates diabetic).

We need to find a good classifier to solve the problem of automatising this task, so we need you to assess different classifiers to know which one will work better.

2. Sections that your report should contain

In any classification experiment there are some design decisions that you should make. Find some of them below:

- **Initial exploration of the data**: You can see how the diagnoses (i.e. the target classes) are distributed, the ranges of each value, etc.
- **Division of the dataset**: The experiment can be carried out by dividing the provided dataset by means of holdout (i.e., training-test), training-validation-test, cross validation, etc.
  If the division is carried out by means of holdout, it is a good practice to **repeat the experiment several times** and report the **average and std of the obtained results**.
- **Data preprocessing**: In previous labs, you have seen that the data was normalized so that the training set had mean 0 and standard deviation 1, and the test set was normalized accordingly, using the StandardScaler class. Once it is instantiated, the StandardScaler is fit with the training data and then used to transform the training and the test sets separately.
  This can be useful when the ranges any variable (column) is very different from the others.
  Even though in some datasets the results are better when the data is normalized, it is not always the case, so **it is useful to assess whether it is better to normalize or not**.
- **Classifier to use**: A classifier that shows a good performance with a dataset may not necessarily work well with other datasets. Therefore, it is important to assess different classifiers, each one with an appropriate tuning of its parameters.
- **Performance metric**: Accuracy is not always the best choice to report the performance of a classifier, especially when the dataset is very imbalanced. There are other cost-sensitive metrics, such as precision, recall, F1-score, sensitivity, specificity, Confusion matrices, ROC curves, etc. An important decision, therefore, is what performance metric (or metrics) to report.
  Sometimes, a performance metric has high values, whether others have poor values, as it has been studied in theory. You must check the results and, if it happens, discuss why.

Besides making these decisions when you make the experiment, you must **reflect them in your report**. In a scientific in which machine learning methods are assessed for a given problem, you will often find the following Sections (or analogous section organizations). **The Sections that you must complete for this assignment are highlighted in yellow**:

Abstract: First of all, a summary of the paper is provided, in a few words (sometimes, not more than 250).

1. **Introduction**
   This Section introduces the problem (e.g., problem that the authors want to solve, why the method they present is important, etc.)
2. **Literature review**

Recent papers covering the same or similar problems are cited, presented, and discussed. This Section tries to establish a context in which where the contribution with respect to previous works is placed.

3. **Methodology**
   This Section must show a description/explanation of the proposed methods.
4. **Experiments**
   a. **Dataset**
      What is the dataset that has been used and its characteristics. Some characteristics you can comment are: the number of variables, number of samples, distribution of the classes (i.e., the number of samples per class), etc.

   b. **Experimental setup**
      This subsection explains **how the experiments have been conducted**. For example: the dataset division, if the data has been preprocessed and how, what are the assessed classifiers and the parameters that have been tuned for all of them, etc.. If more than one experiment has been done, this subsection can be divided.
5. **Results and discussion**
   In this Section, the results are presented. Also, they are commented, and a brief discussion is provided. For instance, if a classifier shows an unexpected performance, or if different performance metrics show values which are very different, it must be commented, and possible explanations should be given. If more than one experiment has been done, this Section can be divided.
6. **Conclusion**
   In this Section, a summary of the paper is provided, including with the most remarkable findings.
7. **Bibliography**
   The full citation of the commented papers (e.g. in the Literature review) is provided.

Of course, in this lab, you will not have to write all these Sections. You are asked only to find the best possible classifier for the problem commented in Section 1.1, so in the report you only need to write the information that would only be contained in Sections 4 and 5 (highlighted).

A document with a length of 2-3 pages should be enough to describe the experimental setup and the results.

A good idea to show the results is using tables. It makes them easier to see and compare them. See an example for the classifier SVM below:

C	Kernel	Kernel parameters	Accuracy	Precision	Recall	...
1	Lineal	n/a				...
	RBF	gamma=auto				...
		gamma=0.25				
		gamma=0.5				

5	Lineal	n/a				
	RBF	gamma=auto				
		gamma=0.25				
		gamma=0.5				
...	...	...	...	...	...	...

3. Requirements for correct assessment

Find below the assessment criteria that will be used to qualify the lab:

Criteria	**Mark**
More than three classifiers are compared with different parameters each one, are evaluated. Using classifiers that have not been used in previous labs is assessed positively.	25%
Different performance metrics are used and reported	10%
The results are shown and discussed clearly, concisely and ordered.	25%
Clear and complete explanations about how the experiments have been carried out are provided.	20%
The delivered code allows to replicate the experiments	15%

In summary, you must deliver:

- **A report** describing some details of the dataset, the experimental setup and the results (remember, 2-3 pages should be enough).
- The **code** that allows to make the experiments.