Aprendizagem Automática

Mestrado em Engenharia Informática

Assignment - step 5

1 Objectives

The **general objective** of this assignment is to apply the different machine learning methods on a dataset, analyze and understand the results obtained.

2 Task

The dataset to be used in the assignment is Urbansound8k [1]. It consists of 8732 labeled audio files approximately 4 seconds long (Table 1). Each recording is labeled in one of 10 classes: air_conditiones, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren and street_music.

Class	# Examples
Air Conditioner	974
Car Horn	429
Children Playing	1000
Dog Bark	999
Drilling	978
Engine Idling	1000
Gun Shot	374
Jackhammer	1000
Siren	920
Street Music	1027

Tabela 1: Class Distribution.

What will be made available for you to use in building your models will not be the original audios from the dataset, but rather a set of features that were collected from these audios. The features that were collected and will be made available are MFCCs. MFCCs are commonly used in sound analysis, such as ambient sound or speech recognition. You can learn more about how MFCCs are calculated in [2].

From the MFCCs collection, you will have available, for each sound, 13 means and 13 standard deviations. The features will be made available in csv files, one for each class, where each line of the csv file corresponds to the features of each sound.

For the first tasks proposed, **each group must choose only three classes**, among the 10 possible. **Each group must work with a different set of classes**.

After choosing the data, and before starting to use it, carry out an exploratory analysis to obtain more information from the dataset:

- Descriptive Statistics
- Univariate Analysis (Distribution of individual features)
- Bivariate Analysis (Correlation between features and the different target variables)

What relevant information can you extract from the Univariate and Bivariate Analysis?

3 Methods Application

Consider using the following methods: Logistic Regression, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Applying the methods to the chosen data, try to decide which method is most appropriate for the problem, giving reasons for your choice. Use the following resampling methods for the various suggested models:

- Holdout
- Cross Validation (with $k = 5$ and $k = 10$)
- Leave One Out Cross Validation (LOOCV)
- Bootstrap

Use the evaluation metrics that you find most appropriate to evaluate the results obtained in each experiment. Analyzing the results obtained, indicate how the variance is affected by the resampling methods used.

4 Feature Selection

Can classification models obtain better results if they use just a few features instead of all available features? Evaluate this hypothesis, using regularization methods.

5 Learning a non-linear function

Use Generalized Additive Models (GAMs) to perform binary classification of your dataset. To do this, you must try to build a model that allows you, among the three classes under analysis, to identify one of them. You should test the three hypotheses and present only the one with the best results. To validate the performance of the models, use cross-validation. Evaluate the results using the evaluation metrics that you consider appropriate.

6 Decision Trees and Random Forest

1. Decision Trees
   (a) Using Decision Trees, build a classification model that allows you to differentiate the classes under analysis.
   (b) Tune the Decision Tree hyperparameters, ensuring that your model is not overfitting the training data.
2. Random Forest
   (a) Using Random Forest, build a classification model that allows you to differentiate the classes under analysis.
   (b) Tune the Random Forest hyperparameters, ensuring that your model is not overfitting the training data.
   (c) After building your Random Forest model, present an ordered list, with the importance of the features used by the model.
   (d) Try to correlate the results obtained in the previous question, with the Univariate and Bivariate analysis carried out in Section 2, and with the results obtained after applying the Ridge and Lasso methods in Section 4.

7 Support Vector Machine (SVM)

Using SVMs, build a classification model that allows you to differentiate the classes under analysis. In this task you must:

- Test all possible kernels;
- Tune the SVM hyperparameters, ensuring that your model is not overfitting the training data;
- Present the SVM model with the best performance on your data, justifying the choice (you should use results from models used in previous tasks to justify your answer).

8 Principal component analysis (PCA)

Use the PCA method to perform feature selection in your dataset. Using the result of the feature selection performed with PCA, evaluate whether the models used previously can achieve better performance.

Compare the results obtained with those obtained in previous tasks (especially with the results from 4 - Feature Selection). What can you conclude about feature selection using PCA?

9 Submissions

A notebook with answers to the proposed tasks. The notebook is .ipynb by default. Any other format must be easily readable. Please take care with the following:

- Steps taken must be succinctly described (through comments in the code or text cells in the notebook)
- Results must be summarized as much as possible.

9.1 Groups

- Assignments are submitted by groups of 2 or 3 students. Different elements may have different grades based on the contribution distribution and interactions about the assignment.
- Code of Conduct
  - All the materials used and consulted must be credited in the work as references.
  - All students should know the Disciplinary Regulations for Students of Polytechnic Institute of Porto (https://dre.pt/dre/detalhe/despacho/4103-2013-2301392)
- It is mandatory the Bitbucket version control tool. Each group must share the repository with PL teacher.

9.2 Deadline

There two mandatory deliveries of the work in Moodle:

- 12th November, intermediate delivery, for feedback
- 17th December, final delivery, for evaluation

Only submissions on the Moodle, before the deadline, will be considered to evaluation. Submissions after that date will not considered. The name of the zip file should be: `APRAU_AAA_CCC_Num1_Num2_Num3.zip`, where: AAA is the teacher´s acronym, CCC the class and Numx the number of each student.

The presentation and discussion, mandatory for all group members, will be on a date to be scheduled by the PL teacher (cf. FUC APRAU course).

10 Bibliography

[1] https://urbansounddataset.weebly.com/urbansound8k.html

[2] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review,"in IEEE Access, vol. 10, pp. 122136-122158, 2022, doi: 10.1109/ACCESS.2022.3223444.