Aprendizagem Automática (APRAU)

Mestrado em Engenharia Informática

Assignment - step 4 (week 9)

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 you'll be working with consists of various environmental and geographical features measured in specific regions. Each instance in the dataset corresponds to a patch of land, and for each one there is an associated Vegetation Type, based on the given attributes.

Features
The dataset contains the following types of features:

- Altitude: The altitude of the terrain in meters.
- Slope_Orientation: The compass direction that the terrain faces (in degrees).
- Slope: Slope in degrees.
- Horizontal_Distance_To_Water: Horizontal Distance to nearest surface water (in meters).
- Vertical_Distance_To_Water: Vertical Distance to nearest surface water (in meters).
- Horizontal_Distance_To_Roadways: Horizontal Distance to nearest roadway, in meters (which could indicate human activity influence on the environment).
- Shadow_Index: Shadow Index measured at three times of the day (9h, 12h and 15h).
- Horizontal_Distance_To_Fire_Points: Horizontal Distance to nearest wildfire ignition points (in meters).
- Canopy_Density: The percentage of land area covered by tree canopies (measured as a percentage from 0 to 100).
- Rainfall_Summer: Average precipitation during the summer (measurement in mm).
- Rainfall_Winter: Average precipitation during the winter (measurement in mm).
- Wind_Exposure_Level: Average wind speed during the winter (measurement in Km/h).
- Soil_Type: Soil Type designation. In the dataset there are 40 different types of soils.
- Wilderness_Area: In the dataset there are 4 different types of Wilderness area.
- Vegetation_Type: The target variable. There are 7 possible classes, each representing a different type of vegetation.

Class	# Examples
Type_1	2160
Type_2	1404
Type_3	1620
Type_4	1080
Type_5	1944
Type_6	2160
Type_7	864

Tabela 1: Class Distribution.

The 7 $Vegetation\_Types$ that we have in the dataset corresponds to different types of ecosystems found in varying geographic and climatic conditions, which play a crucial role in the type of vegetation that grows in each location.

The dataset is imbalanced, meaning that each $Vegetation\_Type$ doesn't have the same amount of examples in the dataset. Table 1 shows the number of examples that each class has.

What will be made available for you to use in building your models will not be the original dataset, but rather a subset of this. The features will be made available in csv files, one for each class, where each line of the csv file corresponds to one sample of ground.

For each group of students, a set of three classes from the dataset will be distributed, among the 7 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.

Note: For the following tasks, you must use the most relevant predictors, based on the results obtained in Task 4 - Feature Selection.

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 identify 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 Reinforcement Learning (RL)

In this task, the goal is to design a system where Q-learning is used to sequentially select features for a classification model (must choose the two best previous models). The idea is to treat the feature selection process as a reinforcement learning problem, where an agent learns to choose which features to use to build the best classification model.

Problem Formulation:

- State: The state is a binary vector representing the selected features so far. For instance, if the dataset has 14 features, the state is a 14-dimensional vector, where 1 means the feature is selected, and 0 means it is not.
- Actions: The action space is the set of all remaining features that haven't been selected yet. The agent can either select a feature or decide to stop selecting features (when it has enough information).
- Rewards:

  - A positive reward is given when the selection of features leads to an improvement in classification accuracy (after training a model with the current set of selected features).

  - A negative reward is given if the selected feature doesn't improve performance or leads to overfitting.

  - A penalty is applied for selecting too many features (to encourage simplicity and avoid overfitting).
- Goal: The agent's goal is to learn to select an optimal subset of features that maximizes the accuracy of the classifier while minimizing the number of features used.

10 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.

10.1 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 Github version control tool. Each group must share the repository with PL teacher.

10.2 Deadline

There **two mandatory deliveries** of the work in Moodle:

- **27th October**, intermediate delivery (25% final grade)
- **29th December**, final delivery (45% final grade)

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).