MSE | MASTER OF SCIENCE IN ENGINEERING

HES-SO

Machine Learning

Practical work 09 – Classification Trees and Random Forests

Teachers: A. Perez-Uribe (Email: andres.perez-uribe@heig-vd.ch) & J. Hennebert
Assistants: Benoît Hohl (Email: benoit.hohl@heig-vd.ch) and Simon Walther (Email: simon.walther@heig-vd.ch)

Summary for the organization:

- Submit the solutions of the practical work before Monday 20.11.23, 23h59 via Moodle.
- Modality: PDF report (max. 8 pages)
- The file name must contain the number of the practical work, followed by the names of the team members by alphabetical order, for example 09_dupont_muller_smith.pdf.
- Put also the name of the team members in the body of the notebook (or report).
- Only one submission per team.

0. Notebooks

Download the notebook material from the Moodle platform

1. Classification Trees

Read the notebook material, follow the instructions, run/modify/complete the code to perform the exercises, and answer the questions.

- Q1.1: At which frequencies does the electrical activity mostly occur?
- Q1.2: Can you easily distinguish between classes visually? What can you say about the inter- vs intra-class variability?
- Q1.3: Describe both of these criteria. What does a gini impurity of 0 means? What does an entropy of 1 mean?
- Q1.4: This problem suffers from a common issue in machine learning. What is this problem called? What could be its causes? How can it be resolved?

- Q1.5: Use the visualization of this tree to show and explain:
    - What is a node? What is an edge? What is a leaf?
    - What are the two additional hyperparameters doing? Do you think that both are necessary in this particular case (min_samples_leaf = 20, max_depth= 4)? Why?
    - What does the color of each node represent?
- Q1.6: Choose one of the nodes. Explain precisely the information given on each line of text in this node.
- Q1.7: Does model 2 still have the same problem as model 1? Explain based on the classification reports and the confusion matrices.
- Q1.8: One of the class seems more difficult to predict than others? Which one? Where could this difficulty come from in your opinion?
- Q1.9: What does this hyperparameter do? Explain giving examples from this dataset.
- Q1.10: Compare results from model 2 and model 3. What are the pros and cons of each of them?

2. Random Forests

Read the notebook material, follow the instructions, run/modify/complete the code to perform the exercises, and answer the questions.

- Q2.1: For each of the hyperparameter: Is there a range of value giving particularly good results? Or particularly bad results?
- Q2.2: These representations give valuable information about hyperparameters. It is nevertheless insufficient. What are/is the main problem(s) with those graphs in your opinion?
- Q2.3: What do the following plots represent?
- Q2.4: What do the white spots (=empty spots) in the heatmaps mean?
- Q2.5: How do those plots address the limitations of the previous visualizations?
- Q2.6: What is grid search? Explain by giving real examples from this specific task.
- Q2.7: Use the plots above to narrow the range of hyperparameters you want to explore. Which values did you choose to test for each parameter? Justify your choices.
- Exercise 2.1: Once you know which values of hyperparameters you want to explore, complete the following code to perform a grid search on those values. Remember that the more values you choose to test, the longer the computational time required.
- Exercise 2.2: Complete the code in order to choose one final value for each of the hyperparameters and train the model.

- Q2.8: Which value did you choose for each hyperparameter? Why?
- Q2.10: The test set should be used only at this stage, and it is theoretically important not to change the hyperparameters based on the performance on the test set. Why?
- Q2.11: Comment your results. -> How well does the model generalize on unseen data? Is a random forest better than a single classification tree in this case? What is the main challenge of this dataset? …
- Q2.11: How is this importance calculate?
- Q2.12: What can you conclude from this graph?

3. Gradient Boosting for classification

Read the notebook material, follow the instructions, run/modify/complete the code to perform the exercises, and answer the questions.

- Q3.1: Two additional hyperparameters were added compared to the RandomForestClassifier. What are these hyperparameters, and what roles do they play?
- Q3.2: Comment the results. Compare these results with the ones obtained with the RandomForestClassifier. Compare more specifically the precision, the recall and the f1-score of the 'r' class obtained with GradientBoostingClassifier and RandomForestClassifier. What are your conclusions?

Summary of work to include in the report

- No need to write a “scientific” report for this PW. Just answer to the questions above as clearly and concisely as possible.
- For exercise 1.1, add the plots to the report.
- For question 1.4, you might add the image of the tree in the report to answer to the question.
- You can add other plots or screenshots to improve the clarity of your answers if you wish.
- Best of luck with your work! ☺

Comments:

Q1.5: min_samples_leaf=20 might not be necessary as each leaf has a number of samples much higher than 20 anyway.

Q2.2: Le problème principal est que ces illustrations ne permettent pas de voir les intéractions entre les différents hyperparamètres

Q3.2: bien

Général: bon rapport

Grade: Pass