Machine Learning Project

Dr. Michelle Lochner

Instructions

- Pracs must be submitted using iKamva in the form of a pdf file.
- You may use any word processing program you like to write your reports but they must be saved as pdf.
- Your file must be named with your student number, followed by `_prac_ml` For example, if I were to submit this weeks' prac it would be called `1234567_prac_ml.pdf`.
- Pracs may be submitted after the deadline, late submissions are penalised as per our late assignments policy.
- Along with your pdf, you must submit a jupyter notebook with all your code. It must be renamed with your student number first, just like the pdf (for example `1234567_prac_ml.ipynb`).
- Your pdf is the only thing that will be marked, the jupyter notebook is just for reference in case we need to run your code. Make sure all code, comments and outputs/question answers are all included in the pdf!
- You must complete the feedback form about the prac before the submission deadline (link at the end of the prac).
- All code must be heavily commented, you should aim to comment every line.
- IMPORTANT: if you take code from an online resource, you must correctly reference the source, including a direct link to the website. You must also include a short statement about what parts you copied from the source and what you changed. Ensure each line of code is commented to show your understanding. Failing to correctly reference your sources could be flagged as plagiarism.

Question 1: Classification algorithm

20 marks (2,10,6,6,6)

For this project, you will complete the notebook in `tutorial-supernovae.ipynb`. Each student will use a different machine learning algorithm. These are all available in the package `scikit-learn`. The algorithms are as follows:

Student name	**Algorithm**	**Python import statement**
Neo Mohlomi	k-Nearest Neighbours	from sklearn.neighbors import KNeighborsClassifier
Siyabulela Sokweleti	Random Forest	from sklearn.ensemble import RandomForestClassifier
Topollo Naketsana	Support Vector Machines	from sklearn.svm import SVC
Alungile Zondo	Logistic Regression	from sklearn.linear_model import LogisticRegression
Mnqobi Zuma	Gradient Boosting	from sklearn.ensemble import GradientBoostingClassifier
Moorane Makwela	Neural Network (MLP)	from sklearn.neural_network import MLPClassifier

Write a description of how your algorithm works. This should be at the level of introductory material for a thesis. Include relevant references with an appropriate referencing style. You description should be approximately 1-2 pages. Be sure to cover the following points:

1. Correct reference to the paper that first introduced the algorithm.
2. Full description of how the algorithm works.
3. Description of relevant hyperparameters.
4. Advantages and disadvantages of the algorithm.
5. Examples of where the algorithm has been used in the literature.

Question 2: Classification choices

10 marks (4,3,3)

You need to run a machine learning algorithm to classify different supernovae. Here you will need to state and motivate the choices you made in the tutorial notebook.

Questions:

1. What's a reasonable choice here for how much data should go into your test set? What are you going to do to ensure you don't overfit?
2. Does your algorithm require the features to be rescaled? Why or why not?
3. Did you use the default hyperparameter values? If not, how did you select them?

Question 3: Classification performance

30 marks (10,5,5,5,5)

Now it's time to assess the performance of your algorithm.

Questions:

1. Include your code for running the classifer in your report. Make sure to include reading the data, splitting into training and test sets, and any hyperparameter selection.
2. Make a confusion matrix plot for your test set. Include the code and plot in your report.
3. Report your accuracy, precision and recall from your test set. Make sure to include your code for computing these metrics (don't use the built in functions, write your own to make sure you understand them).
4. Make a ROC curve plot, where type Ia supernovae are considered the positive class. Include your code, the plot and report the AUC.
5. How good is this performance? Could we use this algorithm to classify supernovae and use them for cosmology? Are there any risks to using your classifier for this?

Have you followed the instructions correctly?

Check each of the following to make sure you've completed them correctly. If not, you will **lose** the marks indicated in brackets (up to a maximum of 10 marks).

- □ *(2 marks)* Answers submitted in the form of a PDF report
- □ *(2 marks)* Code is formatted correctly in the PDF (see video on iKamva for guidance)
- □ *(2 marks)* PDF is named with your student number first like this: `1234567_prac_ml.pdf`
- □ *(2 marks)* Separate jupyter notebook also submitted, containing all your code, and also named with your student number like this: `1234567_prac_ml.ipynb`
- □ *(2 marks)* Feedback form filled in: https://forms.gle/oc6HAU1dkPPsmSbn9