Assessment description

Data Engineers are often required to undertake supervised learning, (classification) in many real-world projects on a daily basis. This task will help prepare you for this scenario by building on theoretical knowledge introduced in the workshops and giving you a deeper understanding of how to apply machine learning algorithms to solve real-world problems. In particular, this task will help you by:

Giving you an opportunity to apply theory into practice
Exposing you to a real-life scenario
Enhancing your understanding of theoretical concepts introduced in workshops (for instance, the definition of each problem, the logic of the algorithms, the meaning of their parameters and hyperparameters and the pros and cons of each).
Building up your problem-solving skills with coding practice
Developing your documentation skills with report output
Enhancing your collaboration skills
Giving you feedback so that you can correct conceptual misunderstandings as you move forward
Assessment details

Reminder: Always refer to the textbook 'hands-on machine learning with Scikit-Learn, Keras & TensorFlow' for coding help.

This assessment will focus on applying two Multiclass classification algorithms to the MNIST datasets.

Use the following steps to complete this task:

Step 1: Split the whole dataset into training/testing dataset either by yourself or by functions provided by python scikit-learn packages. Write an explanation of how you do the split. (5%)
Step 2: Exploit the performance of kNN classifier on the dataset (15%)

Set k=10 which is the ground-truth number of categories in the dataset, then apply the kNN classifier (using functions from Python package). Write a detailed explanation of the meaning of the parameters for the function interface you choose. (5%)
Evaluate the performance with the metrics introduced in the week 5 topic workshops, in particular, the evaluation metrics topic workshop (also using the metric functions provided by Python scikit-learn package). Write a detailed explanation of why a specific evaluation metric is picked and what you can tell from the results about the model. (5%)
Experiment with different values of k, and compare the performance. What did you discover from the comparison? Can you give any reasons for choosing k? (5%)
Step 3: Exploit the performance of SVM classifier on the dataset (15%)
Try appropriate SVM classifiers from the support vector machines modules. Write a detailed explanation of the meaning of the parameters of the function interface you choose. (10%)
Write down a 200-300 word comparison of the performance of the SVM classifier and the kNN classifier. Be sure to include recommendations about the performance of each, and which model you would use in different situations in the future. (5%)
Step 4: Record a presentation video of your code (record your screen in Zoom or similar software) and take us through your code. Use succinct language to explain what you've done and what you've discovered in this assessment in 5 minutes. (5%)
Step 5: Bonus: fine-tune model with grid search (1%)

Apply grid search for either of the classifiers
Summarize/Report your insights in 200-300 words.
Step 6: Bonus: Better evaluation with cross-validation (1%)

Apply cross-validation to evaluate either of the classifiers