COMP10200: Assignment 2 – Part 2

Overview

For this assignment, you will obtain some machine learning data for classification, test several versions of the K-Nearest Neighbour algorithm and a Decision Tree learner using that data, and then write a short report on your findings.

This is Part 2, in which you use the sklearn implementation of Decision Trees on the data set from part 1.

The Data

You must use the same data set as Part 1.

The Code

The code you use for this part of the assignment should be written in Python and should take maximum advantage of the tools in sklearn and numpy.

The Task

Your task is to test at least 8 different versions of the sklearn Decision Tree algorithm to see how they perform on your data set. Create these 8 different versions by choosing 2 values for each of 3 parameters and then trying all the permutations (e.g. vary the criterion, max_depth, max_leaf_nodes, min_samples_split, min_samples_leaf, or others – see https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html). Feel free to try more than 8 versions.

As in part a, you should report averaged results over many runs of the algorithm for each version, each with a different training/testing split. Because decision trees run much faster than kNN, you should easily be able to do 50 runs or more for each version.

The Report

You should expand on your report for part 1. Imagine the boss asked you to test both k-NN and Decision Tree and come up with an overall recommendation. You can expand the sections of the kNN report to also discuss your Decision Tree results, as follows

1. **Data Set Description** – you don't have to add anything here, since it's already done for Part 1
2. **Description of the Tests** (describe the parameter settings you experimented with, describe how you approached generating different training/testing data for each run, etc.)
3. **Results** (accuracy for at least 5 runs for all 8 versions, plus average accuracy for each version, and at least one graph summarizing your results. **Include a graphical view of at least one decision tree** that was produced for one of your better runs – you should be able to drag the pdf produced by graphviz into your report and resize it to fit the page.)
4. **Discussion** (are there clear winners or losers in the versions you tried? Which of the 3 variations of decision trees seemed to make the most difference? How did the Decision Trees perform vs.

k-NN? Give some solid ideas for why some versions were better and why Decision trees or k-NN were better overall. Be as specific as you can and reference the properties of your data. End with a recommendation for your boss – do you recommend k-NN or Decision Trees, and which configuration of the recommended algorithm would be best?)

5. **Future Work** (If you had more time, what more could you explore?)

Throughout your report, make sure you are using standard grammar and spelling, and make sure you make proper use of correct machine learning technology wherever appropriate.

If you quote or reference any information about Decision Trees, k-NN, or issues in Machine Learning that were not explicitly covered in class, you should cite the source for that information using correct APA format.

Feel free to use the report template on Canvas to structure your report.

Report Option: A Video Presentation

If you would prefer, you could also record a video presentation of your results and submit that instead, along with a handout or slides show the results. Your video should include all the information that would be in the report (see items 1 through 5 above). The handout or slides should show all the information from part 3 above (you can use that part of the report template) and any references for quotes or info that you used that wasn't covered in class (in APA format). When you get to part 3 of the report, you can just refer the viewer to the handout or slides, you don't have to read them out.

The content of your presentation will be judged using the same rubric as the report would be, just replace the phrase "well written" with "well presented". Make sure you're using correct machine learning technology wherever appropriate.

Note: If you are taking the video presentation option, you should probably wait until both parts of the assignment are complete before recording, then combine the two presentations into one.

Other Ideas (Optional)

k-NN in SKLearn

Try the SKLearn implementation of k-NN and see how it stacks up against yours. Do your results more-or-less agree? Why or why not?

Feature Selection

You might notice that some features are used a lot more often than others in the decision trees produced by your code. In fact, there might be some features that are not used at all. Perhaps you can use this feature selection to improve k-NN. For example, you could weight some features more heavily than others in the distance calculation. Maybe features that don't appear in your trees should be given zero weight. Can you improve on k-NN in this way?

Handing In

Wait until you have completed both parts of the assignment before handing in. Then zip up your report, your data file(s), and the code for all the versions of kNN and Decision Trees you used. It should be possible for me to unzip your folder and run your code easily (i.e. without having to move files around or

make changes to the code). If necessary, include instructions at the top of each code file with the correct command to run the code.

See the drop box for the exact due date.

Evaluation

This assignment will be evaluated based on: 1. the quality of the report you produce; 2. how well you met the requirements of the assignment; and 3. the quality of the code you handed in (including quality of documentation and referencing within the code).

See the rubric in the drop box for more information.