COMP10200: Assignment 3 – Part 1

Overview

For this assignment, you will obtain some machine learning data for classification, test various SKLearn implementations of Naïve Bayes, and report your findings.

This is Part 1, in which you will use Gaussian Naïve Bayes on your data set from Assignment 2.

The Data & Code

You can use the same data set as Assignment 2, or you can pick a different data set with numeric features.

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 the sklearn Gaussian Naïve Bayes algorithm to see how it performs on your data set. There should be many different runs (at least 50), each with a different training/testing split. See Canvas and the SKLearn documentation for sample code and help:

> https://scikit-learn.org/stable/modules/naive_bayes.html
>
> https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

You should also look at the probabilities generated by the classifier to get a sense of its confidence in its answers. You should compute its average probability score for its correct and incorrect predictions.

The Report

You should write a very short report for this part of the assignment, using word processing software, that contains the following information: The name and link to the data set, the average accuracy and the average probability score for both correct and incorrect predictions. Then compare the performance of Naïve Bayes to your reported performance of kNN and Decision Trees and discuss why Naïve Bayes might have done better or worse than the other two algorithms. Finally, discuss the average probability scores and give an example of how you might use the probability score to provide more useful information than simply a classification label.

Other Ideas (Optional)

"Not Sure"

Maybe if the predictions are close (e.g. within 0.1 or 0.2 for the two highest classes) or uncertain (e.g. probability less than 0.8) the classifier should refrain from making a prediction. Try to implement this, and report the percent of the total that it was uncertain about, and the accuracy among those predictions it was certain about. Does this improve the accuracy measure? In what domains do you think it might make sense to have an "uncertain" classification?

Class Weighting

Sklearn's GaussianNB contains an optional parameter for specifying the prior probabilities of each class. If you don't specify, it estimates class probabilities from the data. Experiment with this parameter to see if you can improve performance.

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. 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.