TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Week 4 Assignment CS7CS4/CSU44061 Machine Learning

Rules of the game:

- **Its ok to discuss with others, but do not show any code you write to others. You must write answers in your own words and write code entirely yourself. All submissions will be checked for plagiarism.**
- **Reports must be typed (no handwritten answers please) and submitted as a separate pdf on Blackboard (not as part of a zip file please).**
- **Important: For each problem, your primary aim is to articulate that you understand what you're doing - not just running a program and quoting numbers it outputs. Long rambling answers and "brain dumps" are not the way to achieve this. If you write code to carry out a calculation you need to discuss/explain what that code does, and if you present numerical results you need to discuss their interpretation. Generally most of the credit is given for the explanation/analysis as opposed to the code/numerical answer. Saying "see code" is not good enough, even if code contains comments. Similarly, standalone numbers or plots without further comment is not good enough.**
- **When your answer includes a plot be sure to (i) label the axes, (ii) make sure all the text (including axes labels/ticks) is large enough to be clearly legible and (iii) explain in text what the plot shows.**
- **Include the source of code written for the assignment as an appendix in your submitted pdf report. Also include a separate zip file containing the executable code and any data files needed. Programs should be running code written in Python, and should load data etc when run so that we can unzip your submission and just directly run it to check that it works. Keep code brief and clean with meaningful variable names etc.**
- **Reports should typically be about 5 pages, with 10 pages the upper limit (excluding appendix with code). If you go over 10 pages then the extra pages will not be marked.**
- **When selecting a hyperparameter value show cross-validation analysis to justify your choice.**
- **When evaluating an ML algorithm on data, compare against a baseline model.**

Downloading Dataset

- This assignment uses two datasets contained within one file. Download the file from `https://www.scss.tcd.ie/Doug.Leith/CSU44061/week4.php`. Each dataset within the file starts with a header line which begins with a #, so you can use this to split the file and extract the two datasets into individual files. Important: You must fetch your own copy of the assignment file, do not use the file downloaded by someone else.
- Please cut and paste the header lines (which begin with a #) and include in your submission as they identify the datasets.
- Each dataset consists of three columns of data (plus the first header line). The first two columns are input features and the third column is the ± 1 valued target value.

.

Assignment

In this assignment you'll go through the full ML workflow for each of the datasets you downloaded: feature selection, model selection, model training and evaluation. Not all datasets are useful, e.g. sometimes the data measured fails to capture the important relationships or is just too noisy. You now have the tools to analyse the data to uncover such problems.

(i) Start with the first dataset.

(a) Using sklearn augment the two features in the dataset with polynomial features and train a Logistic Regression classifier with L_2 penalty added to the cost function. Use cross-validation to select (i) the maximum order of polynomial to use and (ii) the weight C given to the penalty in the cost function. Remember you'll also need to select the range of C values to consider and the range of maximum polynomial orders.

Important: It is essential that you present data, including cross-validation plots and plots of model predictions and the training data, and give clear explanations/analysis to justify your choices. Use the ideas/tools you've learned in the module to help you in this. Be sure to include error bars in cross-validation plots. Rememberr to use a performance measure appropriate to classification (so probably not mean square error).

(b) Now train a kNN classifier on the data. Use cross-validation to select k, again presenting data and explanations/analysis to justify your choice. There is no need to augment the features with polynomial features for a kNN classifier, kNN can already capture nonlinear decision boundaries. Again, it is *essential* that you present data and give clear explanations/analysis to justify your choices.

(c) Calculate the confusion matrices for your trained Logistic Regression and kNN classifier. Also calculate the confusion matrix for one or more baseline classifiers of your choice e.g. one that always predicts the most frequent class in the training data and/or one that makes random predictions.

(d) Plot the ROC curves for your trained Logistic Regression and kNN classifiers. Also plot the point(s) oin the ROC plot corresponding to the baseline classifiers. Be sure to include enough points in the ROC curves to allow the detailed shape to be seen.

(e) Using the data from (c) and (d) evaluate and compare the performance of the Logistic Regression, kNN and baseline classifiers. Is one classifier significantly better or worse than the other? How do their ROC curves compare with that of a random classifier. Which classifier, if any, would you recommend be used? Explain the reasoning behind you recommendation.

(ii) Repeat (i) for the second dataset