Assignment 3 – CSC3060 "AIDA"

Assignment is marked out of 100 marks. Assignment is worth 30% of the module assessment.

Deadline: 11pm Friday, 10th January 2020.

This version: 2019-11-19.

Introduction

In this assignment, we will use the features you developed in Assignment 2 to solve classification problems using machine learning. Specifically, you will fit classifiers to your image data, in order to build and evaluate useful models that can predict the class labels for unseen images. This assignment is to be completed individually.

This assignment must be completed in R. You may not use Excel for any calculations in the assignment, or for figures (using Excel to construct tables for your report is OK).

Convenient and commonly used machine learning packages are available for R, such as "class", "caret" and "randomForest". When you use a procedure that has an element of randomness (e.g. creating cross-validation folds) please use the seed value 3060 (your code should give the same results each time it runs). The seed only needs to be set once at the top of each source file.

Section 1 (30 marks)

For this section, you will make use of the dataset you created for Assignment 2. Your feature file **STUDENTNR_features.csv**, which was uploaded as part of your work on Assignment 2, is the starting point for this task. At a minimum, your "features" file should contain the non-custom features specified at the beginning of Assignment 2 (all students should at least have these features, if you do not then please contact the lecturer, quoting your student number). The file should be placed in the main directory for the assignment.

1.1. Using the **verticalness** feature variable only, fit a single logistic regression model to predict the probability of belonging to the "living" category of doodles, using all 160 items. Present the results table for the logistic regression, including the coefficient estimates, the z-scores and associated p-values. Briefly visualise and interpret the results of the logistic regression.

1.2. Using the logistic regression model you calculated above, create a classifier based on a suitable cut-off value. Find the accuracy of the classifier, over the 160 items. Will the accuracy you have found for this model for these 160 items be representative of the accuracy of this model on other doodles you might hypothetically draw, and which were not included in this set of 160?

1.3. Using any 3 features that you think should be useful (justifying your choices, e.g. on the basis of results and visualisations in Assignment 2), use logistic regression to build a classifier that discriminates between the "living" and "non-living" doodle categories. Use 5-fold cross-validation to evaluate the accuracy of your fitted model. Briefly (1-2 sentences) interpret your results.

1.4 Does your model in subsection 1.3 distinguish between the "living" and "nonliving" categories for the 160 images significantly more accurately than a "random" model that just randomly responds "living" 50% of the time and "nonliving" 50% of the time? Perform a suitable statistical analysis using the binomial distribution.

1.5 Analyse the pattern of errors on the test items for the model in part 1.3 by investigating (a) how often instances of each of the 4 living thing doodle types are incorrectly classified as "nonliving" and (b) how often instances of each of the 4 nonliving thing doodle types are incorrectly classified as "living". Are there any interesting patterns in this accuracy data? Give a very brief qualitative discussion (2 sentences) about what features (not amongst the three you selected) might help further improve the performance of the model.

Section 2 (30 marks)

Larger sets of doodle data have been created for you. You can download the data at the URL that will be emailed to you. These data are for your use only and should not be shared with others (to use other people's data will be considered collusion). These data are to be used in Sections 2, 3 & 4.

These data consist of a dataset of 1000 training items for each of the 8 doodle types (8000 training items in total). The features are in the same format as Assignment 2. The custom features have been set to 0. For the training instances there are three files:

- the features of the image, as described in Assignment 2.
- the 2500 pixels of the image (which you can use to calculate custom features that you think may be useful for classification, if you wish).
- a png image of the item (which may be useful for data inspection).

In this section, you are to perform classification with respect to the 8 doodle categories.

2.1. Perform k-nearest-neighbour classification with all odd values of *k* between 1 and 59 (inclusive) on the training set, using the first 8 features in the "*_features.csv" files (note that the first 2 columns are just the label and the index; the features start from the third column). It is recommended that you use `'knn'` from the `'class'` package for this section. Report the accuracy over the training set for each value of k (use all 160 items in this subsection as training data and do not worry in this subsection about overfitting to the training data; i.e., do not use cross-validation).

2.2. Perform k-nearest-neighbour classification with all odd values of k between 1 and 59 (inclusive), using 5-fold cross-validation, using the same 8 features as in 2.1. Report the cross-validated accuracy for each value of k. Create a figure similar to **FIGURE 2.17** of the ISLR text book, showing the classification error rate (or accuracy rate) over the training set and the cross-validated classification error rate (or accuracy rate) for each value of 1/*k*. Briefly interpret the results of 2.1 and 2.2 with reference to your graph.

2.3. Find out how often each doodle type is confused for each other doodle type, on the basis of your cross-validated results from Section 2.2 (using the value of k that gave the best overall cross-validated accuracy). Consider suitable ways of visualising the results.

Section 3 (40 marks)

In this section, you are to perform classification with respect to the 8 doodle categories.

3.1. Perform classification using bagging of decision trees, using the same 8 features as subsection 2.1, and using bags of size {25, 50, 200, 400,800}. Report the accuracy of fitted models using (a) out-

of-bag estimation and (b) 5-fold cross-validation ((a) and (b) here should be done separately). Briefly explain and interpret the results for this set of models.

3.2 Perform classification with random forests using 5-fold cross-validation. Calculate multiple random forest solutions using number of trees between 25 and 400 (increments of 25) and number of predictors considered at each node = {2, 4, 6, 8}. Evaluate the models using cross-validation. Find the combination of tree-number and predictor-number giving the best cross-validated accuracy (this is called a "grid search" of the two hyper-parameters, number of trees and number of predictors). Briefly visualise, explain and interpret the results for this set of models.

3.3 Random forests and cross-validation have an element of randomness, so let's see how variable the accuracy is across different independent runs**.** For the best model in 3.2 (i.e. best values of tree-number and predictor-number) refit the model 20 times, to obtain 20 cross-validated accuracy scores. Report the mean and standard deviation of the accuracies.

3.4 Adding an additional predictor feature to the model might improve accuracy. Similarly, removing a "bad" feature might not really affect accuracy. Choose a feature to remove from the model (provide a justification for your choice, e.g. on the basis of analyses in Assignment 2). For the same tree-number and predictor-number hyper-parameter values as in 3.3, fit this model and calculate cross-validated accuracies 20 times. On the basis of the accuracy data you have collected, is this 7-feature model significantly less accurate than the 8-feature model reported in subsection 3.3?

Assessment criteria and marking process

The most important criteria in marking is the quality and clarity of your report, including the correctness and accuracy of your models (approximately 75% weighting). In your report, you should demonstrate that you understand the methods used in each sub-task. Explain your reasoning, assumptions and steps of the procedures used. You should explain and interpret your results. What are your results telling you? Are the results what you would expect? If you ran into difficulties, explain what they were and the efforts you made to try to overcome them.

Code has a weighting in marking of approximately 15%. Your code should be clear and logically organised, and do what is required, but code efficiency and code sophistication is not important (this assignment does not require complex programming). Logical organization of code includes appropriate use of variables, iteration, functions, etc., rather than repetition of the same steps with "hard-coded" values. If you use freely licenced code, packages, or libraries (which is encouraged), these should be appropriately referenced (e.g. by citing a URL in a comment). The code must be easy to use and the comments must include information about the required steps to replicate the results that you have obtained and are presenting in your report (transparency and replicability are essential in data analysis). Do not upload unnecessary code (e.g. the entire codebase of some third-party library you are using).

Attention to detail and following the assignment instructions accurately will also be considered in marking (approximately 10% weighting). Each sub-task has a precise specification. Make sure you carefully follow the instructions, and use the features specified for each task, and the specified procedures (number of cross-validation folds, seed value, etc). Make sure you upload your deliverable files in the specified formats.

Deliverables

You must submit your assignment online, using Canvas, by 11pm Friday, 10th January 2019.

The online uploaded file must be a ZIP file called **assignment3_STUDENTNR.zip**, containing multiple files and directories. The contents of the zip file are specified below (**bold** text indicates folder names):

- STUDENTNR_assignment3_report.pdf
- STUDENTNR_features.csv
- code
  - section_1.r
  - section_2.r
  - section_3.r

In addition, I will assume that the doodle data that has been assigned to you is in the following folder:

- doodle_data
  - feature_files
  - pixel_files
  - png_files

However, you should ***not*** upload this folder as part of your submission.

The current working directory should be the location of the source file. Your code should use relative paths; i.e. it should read the training data from "../doodle_data".

A RAR file is not a ZIP file. A broken or corrupt ZIP file is not a ZIP file. It is your responsibility to ensure the assignment is uploaded and double-checked before the deadline.

Please use the provided report template for preparing your report (or create an equivalent LaTEX format). Ensure that the header and footer information (student name, student number) is clearly visible on the printout. The word limit for the report is 4000 words (excluding tables and figures).

By submitting this assignment you acknowledge that it is your own work and that you are aware of university regulations regarding academic offences, including (but not restricted to) plagiarism and collusion.

Standard university penalties apply for late submission.