ECON498 Machine Learning I(Fall 2020) Final Exam

Deadline: 12th December 2020 (23:59)

Name: ____________________________

UID: ____________________________

1. (100 points) The dataset ”business_sample.json” contains information about businesses on yelp.com, which includes the name, the location, the review score (stars), review count, and the category of the business......etc.

   (a) Use the dataset to train a model that predicts the review score (stars) of a business. Since this dataset contains the information of the review score, you can use a supervised learning model. However, you can also choose to use an unsupervised learning model (or a combination of both) if you think that is appropriate.

   (b) Explain why your model is good. You can use the confusion matrices or accuracy scores or any other measures (sklearn.metrics is useful, check the documentation). You can also describe how you clean up the dataset. You may also try using other models to do the same prediction and evaluate why your model is better than them. You can compare your model to any number of other models, but I would say you should include at least one comparison in your work. (It is very difficult to argue that your model is good without a comparison, right?). This is suppose to be a report. You need to write your arguments instead of just handing be a bunch of code and numbers.

   (c) I will use a file very very similar (but not identical!) to the file ”business_no_stars_review.json” to test your program. Notice that this file does NOT contain the review score (”stars”) and it does not contain the review count neither. It means that your model should be able to tell what is the review score simply by reading in the name, address, location, attributes, category, and hours. If your model needs to use review count to predict the stars, you need to predict the review count first.

   (d) You do not need to hand in the dataset but you need to hand in your Python code. The name of your python code should be ”runme.py”. It should read a file named ”business_sample.json” and train a model (part (a)). It then prints out measurements that you used (confusion matrice, accuracy scores......etc.) to supplement your answer in part (b). Finally, it should read in a file ”business_no_stars_review.json” and make a prediction of review score(stars) for each buisness inside this file (part (c)). If your code needs to read in any extra files, you need to submit those files and specify the instructions in a README file.

   (e) Therefore, your answer should include 3 files, 1 written report ”report.pdf” (for part (b)), 1 python file ”runme.py”, and 1 README file (for part (a) (c) and (d)).

   (f) Some parts of the dataset can get you more information than just using the raw data. For example, you may analyze the name of the business using text analysis, or use the location of the business to calculate distances.

   (g) There is no perfect model. Try to make a valid argument about why you think your model is good. For example, sometimes it is very difficult to train a model which is able to distinguish ”stars=3.0” and ”stars=3.5”. Sometimes it is good to just truncate the 0.5 and give up distinguishing them. There is no clear answer to whether ”A model which can distinguish 3.0, 3.5, and 4.0 fairly well” is better than ”A model which cannot distinguish 3.0 and 3.5 but can distinguish 3.0 and 4.0 very well” or not. It all depends on how you make an argument to support your analysis.

   (h) I think most of you will choose to use Python. However, if for some reasons you want to use another computer language, please let me know via email.