CSE/ECE 343/543: Machine Learning
Assignment-4

Max Marks: 50 **Due Date**: 11:59PM, Nov. 20, 2018

Instructions

- Keep collaborations at high level discussions. Copying/Plagiarism will be dealt with strictly.
- Start early, solve the problems yourself. Some of these questions may be asked in Quiz/Exams.
- Submission Instructions: Submissions will be through backpack. Create a single *firstname-A4.zip* file containing a report **A4.pdf**, your source folder **A4-src** and theory questions solutions **A4-theory.pdf**. Report all your theory solutions and outputs of all programming questions e.g intrinsic and extrinsic parameters,figures, images etc in **A4.pdf**. List name of all the functions/scripts that you have implemented along with the two line summary in **A4.pdf**. Put all your programming functions/scripts in **A4-src**. You are allowed to use *numpy*, *scipy* and *matplotlib* only, unless specified otherwise. In case of any doubt, initiate a discussion on backpack or drop an email to any of the TA with the subject line [**ML18-A4-Doubt**]. Emails with other subject lines may suffer delays in response.
- Report(A4.pdf) is **required**. 50% of the total points of the programming question will be deducted if the results are not reported in A4.pdf
- Late submission penalty: As per course policy.

PROGRAMMING QUESTIONS

1. (50 points) **Decision Trees and Random Forest.** To minimize its loss, a bank needs a decision rule regarding who to give approval of the loan and who not to. The given dataset contains information about some loan applicants at a bank in the form of 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk. A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles. Use DecisionTreeClassifier and RandomForestClassifier from sklearn to build two model for this classification problem. Please use the German Credit Dataset for the problem.

    a) (*10 points*) Compare the accuracies of the two models and comment on the performance of each, adn their differences.

    b) (*10 points*) List the hyperparameters that you have used in both cases. Justify your choice using plots/graphs/tables and other analysis tools.

    c) (*10 points*) Demonstrate that your models are trained properly, i.e., show evidence (using plots/graphs) that your models are not overfitting or underfitting.

d) (*15 pints*) Using k-fold cross-validation, check if the validation set error variance for the Decision Tree classifier is higher than that of the Random Forest Classifier.

e) (*5 points*) Save the best model to disk. You can serialize the model in any way you want (preferred: *sklearn*'s joblib function to save models as pickled files). Load the saved model in a separate file to predict the results on test data.