Examination Guidelines for Machine Learning, spring 2021, ITI43210-1 21V

June 24, 2021

The examination in the course consists of two projects (65%) and a theory exam (35%). The projects may be done in cooperation with one or two other students whereas the theory exam is individual. The result of each project should be a written report that contains a maximum of 10 A4 pages in 11 point or 12 point font. It should be written according to the standards of a scientific paper with citations and a reference list.

The grading of the first project is based on how well the students have answered the following points. Obviously, the grade will be lower if answers to some points are missing.

- Describe the problem to be solved and your dataset. How was the raw data generated? What applications does the problem have? What predictors, that is attributes, are there and how are they converted to suitable input for your chosen tree learning algorithms?

  If you have many numerical attributes, consider using PCA for preprocessing. If you use PCA, you will need to scale and center each relevant attribute and reduce skewness using Box-Cox transformations.

  Also consider other ways to perform feature extraction and preprocessing.

- What have others done previously with the same or similar data sets? Search for references on `scholar.google.com`.

- The easiest tree tools to use are C5.0 and Cubist. You should start your project using these.

  If you feel ambitious, you may then consider using random forest (rf) and gradient boosting (xgboost) accessed through the Caret package in R.

- Are trees or rules best? Try to explain the differences if there are any.

- Interpret the output from your tree tool(s) for the data set. Choose a number of rules and see which are sensible and which are not.

- Try to characterize the generalizing ability of the models generated by the tree software using repeated cross-validation or alternatively a test data set. How sensible to missing attributes or less training data is tree learning in your case?
- Do classifications have differing costs? If that is the case, use a `.costs` file in C5.0. Explain and analyze the result.
- Examine other relevant tree algorithm options for your data set, for example winnowing, boosting or pruning. Describe each option.
- Evaluate how good results you have obtained. It is more important to give a correct evaluation and work systematically than to obtain the lowest possible error percentage. What future improvements are there?

The grading of the second project is based on how well the students have answered the following points. Obviously, the grade will be lower if answers to some points are missing.

- Search for references to previous work where neural nets have been used on the same or similar data as yours. Use `scholar.google.com` or `citeseer.ist.psu.edu`.
- In what ways can your data be coded in order to be suitable for neural nets? Do you view your problem as classification or regression? Which codings do you intend to try?
- You can use the neural net packages in R, the neural network toolbox in Matlab, Theano, Tensorflow or your own code. If you are an ambitious student, you may use several of these options.
- Consider how to split the data into sets for training, validation and testing. The validation set is employed to choose architecture, stopping criteria and other parameters. Do you intend to run cross validation? Plan and describe your experimental methodology.
- If it is feasible with your chosen implementation, try at least three different optimization algorithms / momentum schedules in combination with varying numbers of hidden layers and number of nodes in the these layers and number of epochs. How is overfitting related to the various choices?
- Give an interpretation of the output from the neural net for some selected inputs.
- Do you see signs of local optima? The error valley problem? Vanishing gradients?

- Do classifications have differing costs? How can you train the net with that taken into consideration?
- Calculate a confidence interval for the error ratio and describe what it means for the interpretation of your experimental results.
- Evaluate how good results you have obtained. It is more important to give a correct evaluation and work systematically than to obtain the lowest possible error percentage.
- Criticize and compare neural nets with classification and regression trees according to a number of suitable criteria that you choose yourself.
- What future improvements are there?

The course book is Applied Predictive Modeling by Max Kuhn and Kjell Johnson supplemented with manuals for the software that is used, for example for C5.0, Cubist, R, Matlab and scikit-learn.

In 2021, the theory in the course consists of the following.

1. Preprocessing, including scaling, centering, Box-Cox, feature extraction and PCA.
2. Classification and regression trees. How to construct trees using entropy heuristics.
3. Ensemble algorithms including Random Forest and gradient boosting
4. Introduction to neural nets. Gradient descent and other training algorithms. Different activation functions and error measures. Challenges with neural nets and how to overcome them.
5. Convolutional neural nets.

The grading of the exam is partially based on how well the following has been respected.

Clearly explain how each answer has been produced and give good motivations for all calculations. In order to obtain a good grade on an exam problem, all intermediary steps in the method that you use to solve the problem should be carefully presented.

- **You may not use any machine learning software or anything related to it of any kind for solving or helping you solve the exam problems.**
- **The exam is individual, that is cooperation or help in any form is not allowed.**

Generally, the grading of the exam and the project reports is based on the following criteria.

F An exam or report that does not satisfy the minimum criteria,

E An exam or report that satisfies the minimum criteria,

D A satisfactory exam or report but with low independent creativity and analysis.

C An exam or report that is good in most areas and with independent analysis, evaluation and creativity.

B A very good exam or report with a high degree of independent analysis, evaluation and creativity.

A An excellent and clearly outstanding exam or report with a very high degree of independent analysis, evaluation and creativity.