CSC 496 **Introduction to Machine Learning** Spring 25

# Homework No. 3

**Due April. 6 (11:59 pm)**

**(100 points)**

**Objectives**

1. *Apply various classification algorithms to the movie reviews dataset*
2. *Use k-fold cross validation to identify the parameters that optimize performance (generalization) for each method*
3. *Compare the accuracy and explainability of each method*

**Problem #1**

For this homework, you will apply the following classification methods to the *movie reviews classification data* (available in Moodle)

1. KNN (vary the distance measure and K)
2. Multinomial Naïve Bayes
3. Random Forest
4. Gradient Boosted Regression Trees

* Apply 4-fold cross-validation to the provided training data subset to train your classifiers and identify their *optimal parameters*.
* You need to apply each algorithm to the original data and to the data normalized using at least 2 different methods. For each algorithm identify the best normalization (or no normalization) and justify your choice.
* After fixing the classifiers’ parameters, apply each method to the provided testing data subset to predict and analyze your results. *Compare the accuracy* obtained during training (average of the cross-validation folds) to those of the test data and comment on the results (overfitting, underfitting, etc.)
* Analyze the results of each method by *inspecting the feature importance* (if applicable) and few misclassified samples.
* Select the best algorithm and justify your choice based on *accuracy*, *explainability*, *time required to train/test*, etc.

**What to submit?**

* Report: a **single** file (MS Word, PPT, or PDF).
  + **Describes** your experiments,
  + **Summarizes**, **explains** (using concepts covered in lectures) and **compares** the results (using plots, tables, figures)
  + Identifies the best method.
  + The report cannot exceed 10 pages using a font of 12
  + Assign numbers to all your figures/tables/plots and use these numbers to reference them in your discussion
* Audio Recording: does not exceed 10 minutes
  + You need to synchronize your recording with the text/figure that you will be explaining
  + In the first 3 minutes of the recording, you need to show that your code is working and generating output (you run it while recording)
  + In the remaining 7 min, you need to explain your results (figures), compare/explain results, etc.
* Do not submit your source code
* Do not submit raw output generated by your code!