CSE 546 **Introduction to Machine Learning** Fall 2024

# Homework No. 2

**Due Sept. 18 (11:59 pm), 2024**

**(100 points)**

**Objectives**

1. *Build and analyze simple classification algorithms based on KNN and linear models*
2. *Use* ***k-fold cross validation*** *(k=5) to identify the parameters that optimize performance (generalization) for each method*
3. *Identify cases of underfitting and overfitting*
4. *Select parameters that optimize performance (generalization)*
5. *Compare the accuracy and explainability of each method*

**Problem #1**

For this homework, you will apply the following classification methods to the *SPAM e-mail data* (available in Blackboard)

1. KNN binary classifier. Vary the parameter K
2. Logistic Regression classifier. Vary the regularization parameter C
3. Linear Support Vector Machines classifier. Vary the regularization parameter C

* Apply 5-fold cross-validation to the provided training data to train your classifiers and identify their *optimal parameters*.
* After fixing the classifiers’ parameters, apply each method to the provided testing data 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 and identify/explain few (if any) common and different 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!