

# CSE471: Statistical Methods in AI

## Monsoon 2015

**Assignment #1:** K Nearest Neighbor Classifier

**Due:** Before 5:00pm on 25/08/2015

### General Instructions:

- Assignment can be implemented in Matlab/Octave, Python, C/C++, R .
- Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious actions including an **F grade in the course** is possible.
- A single pdf file needs to be uploaded to the Courses Portal. The file should contain your answers as well as the code you have written and its output.
- Include the assignment number, your name and roll number at the top-left of the first page of your submission.
- Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

### Assignment Preamble:

Implement the K-nearest neighbor (KNN) classifier and test it on **any three** different datasets from the UCI Machine learning repository (<http://archive.ics.uci.edu/ml/>) and choose the **fourth** one from any publicly available dataset source.

### Experimental Procedure:

1. Download the datasets.
2. Compute 2 to 5 fold cross validation results and report mean and standard deviation of the accuracy (total # correct test instances/total # of test instances) across each fold.
3. Use K=1,2,3,4,5 in K nearest neighbor classification of the testing set samples in each fold.
4. Handle any ties appropriately but do mention the procedure in the report.

### Include in your report:

1. Please list the details (Number of features, Number of instances, Number of Classes, etc) of the datasets.
2. List out and discuss the choice of distance function used for each dataset.
3. Visualize the results with curves in a 2D plot with y axis being the Mean accuracy and x-axis being the K values in KNN. Add the standard deviation results to the same plot using the error bar (e.g., <http://in.mathworks.com/help/matlab/ref/errorbar.html> ). Include four such plots for each dataset (one plot/fold).
4. Summarize your observations related to the results.