CS-482/682 Machine Learning

K-NN Classifier with Cross Validation

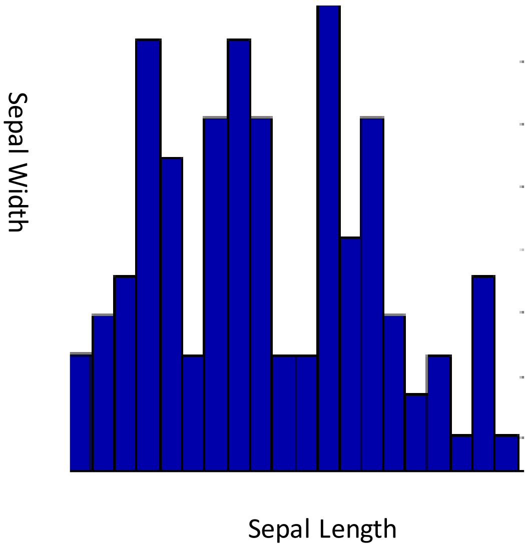
In this assignment, you will demonstrate your knowledge of supervised learning and cross validation techniques by completing the following tasks and writing a report. Note that report can have tables, plots and descriptions. It should not have any code or direct copy of the output from the code. Code should be submitted as a separate file.

1. **Selecting the Dataset:** Use one of the following options for dataset
   a) Wine data set that is classification problem with 3 classes https://archive.ics.uci.edu/ml/datasets/Wine
   b) Breast Cancer Dataset with 2 classes https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
   c) Use a dataset of your choice. Your dataset must be for classification problem with real/integers features only; it must have between 100 to 3000 (`n`) rows. Data must have no missing values and it must not contain any text or categorical data. Note that datasets from UCI or KDNuggets websites are most well studied and popular.

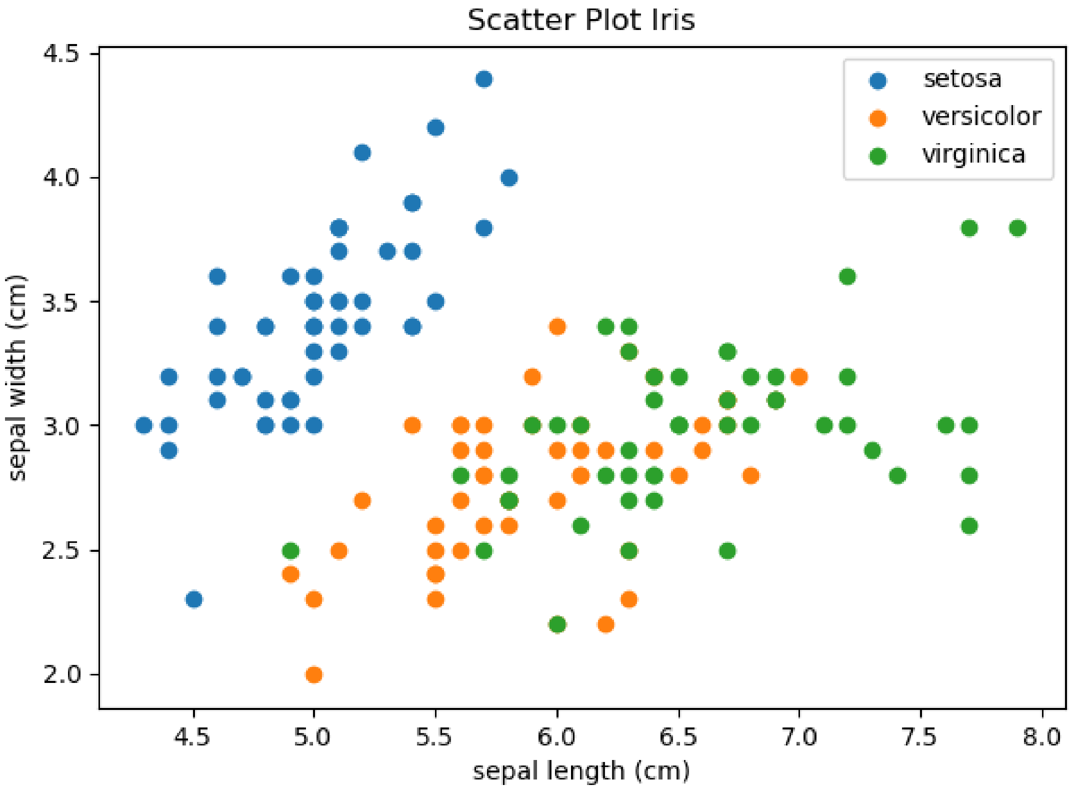
2. **Loading the data**: Write your own load function to read a `csv` file with headers and return the following

   data - numpy array of shape (`n,p+1`)
   target – numpy array of shape (n,)
   feature_names : list of strings
   target name: str
   filename: str with name of csv file read.
   You will use this function in future assignments.

3. **Meet the Data Section:** Provide the following information about the data. This information must be obtained by coding and then included here in the format of a table.
   a) Number of features
   b) Number of samples
   c) Description of features
   d) Description of target
   e) First five rows of the data
   f) Pair Plot Histograms: Manually select 2 most influential features and display using matplotlib a histogram indicating how many samples contain a particular value of the feature. Label your plots. Use blue color for your histogram. A sample is shown below for iris dataset with two selected features of Sepal Length and Sepal Width.



   g) Pair Plot Target: Select 2 influential features and display scatter plot of the target values against these two features. Use different colors for each class of the target with a legend describing what each color represents.



4. **Model Development and Training:** Split the data into 80% training and 20% as test values. Perform k-NN algorithm with various number of neighbors varying from 1 to sqrt(n)+ 3, where n is the number of samples in your dataset. You must skip even numbers in your check such due to possible tie if your problem is binary classification. If it is not binary classification, you can have even numbers as well. *Show the plot similar to the one below and state the value of k that gives the best test accuracy*. Do not use cross validation at this time

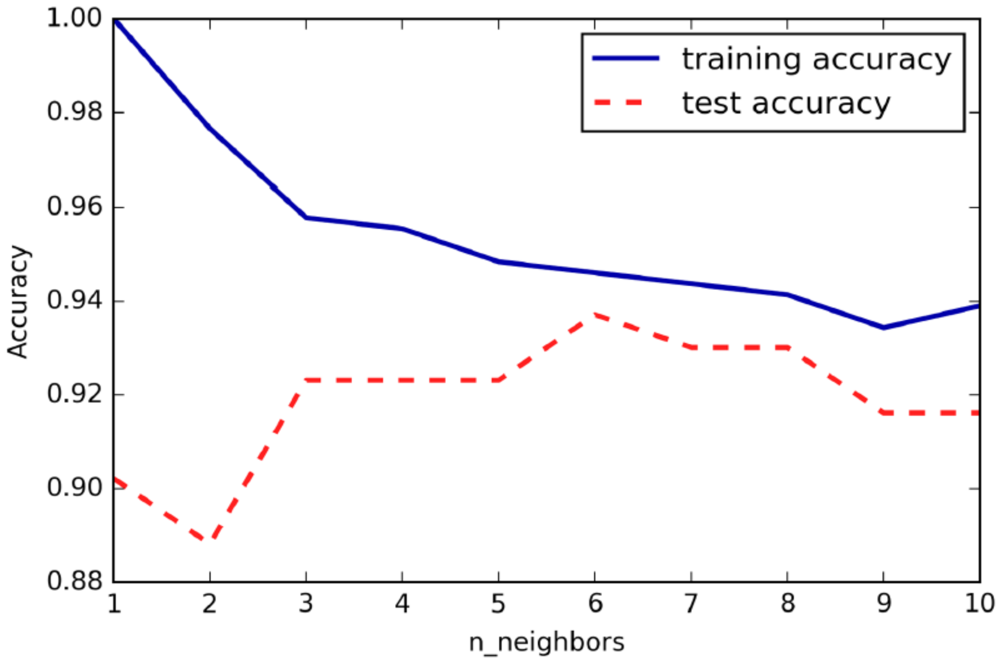


Figure 2-7. Comparison of training and test accuracy as a function of n_neighbors

Best k value is 6.

5. **k-NN with Cross Validation:** Now using the best value of k found in the above step, use 5-fold cross validation with StratifiedKFold. This must also use 20% of the data as test set. Present a table of training and test values for each fold and present the mean accuracy as shown in the sample below. Is the training and test accuracy in Step 4 validated using cross validation? Why or why not?

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Training Accuracy						
Test Accuracy						