Practical 6:
(for Lecture 6: Classification)

Q1. K-NN

Check out the provided program on k-NN, applied to the famous Iris Data set Two parameters are interesting: (i) the ***split*** which is the size of the training subset v test subset (split = .67) means roughly 2/3rds training, 1/3rd testing, (ii) ***k*** which is the size of the collection of nearest neighbours used for the prediction. Note, the ***Accuracy*** of model is determined for test sets; it measures how well it classifications work for this unseen instances, that have been separated from the training set.

Taking ideas about cross-validation into account you should to two things: (i) systematically vary the size of the ***split*** ; exploring five other values for it >0.0 and <0.9 (to be chosen by you), (ii) systematically vary ***k*** on 10 selected values between 1 and 50.

In two separate graphs plot the accuracy scores you get for these these parameter changes.

For each of these graphs , discuss the results found; explain why the graph goes_up/goes_down/is_unchanging when the parameter is varied.

Q2. Bayes Classifiers

Have a look at the **nltk** Bayes Classifier that does the prediction of male/female names based on the last letter in the name. Think of a new feature that you could extract from the data-set; define a method for it (modifying **`gender_features`**). Discuss the results of this change, what new outputs occur in the classification and what things remain the same? [Hint: there is a built-in method **show_most_informative_features** which may be of help; read about it in **nltk**].

Now, find the accuracy score for the original classifier using the last_letter feature. Also, find the accuracy of the classifier using the feature you defined. Discuss why one accuracy score is the same/different to the other score, illustrating your answer with examples of the classifications found.