Shiv Nadar University
CSD361: Introduction to Machine Learning

Assignment #1

(Bayes Decision Rule)

21-1-2022

Max marks: 100
Due: 30-1-2022, 11.59pm

All programs to be written in Python. Submit the Jupyter notebook/lab file on BB.

1. You have to use the Iris data set for building a BDR based classifier. Download the data set from `https://archive.ics.uci.edu/ml/machine-learning-databases/iris/`.

   Use 10% of the data set (randomly chosen) for testing your classifier and the remaining 90% (learning set) to create your classifier. You have to construct a class conditional probability distribution for the 3 classes using the learning set. Set up a suitable sized grid in the feature space (4-dimensional). For each grid volume element give a label based on the BDR. This is your model assuming that the prior probability for each class is equal. When you get a test iris (from the test set) first find the grid element in feature space where it is located, then give it the label of that grid element. Find the error rate on your test set.

   Repeat the above a number of times (say 10) with different random seeds (so that you get different test sets) and find the average error rate. Compare it with the error rates in the literature (you can find it by doing a search on the internet). They are in the region of 2-3%.

   Experiment with different grid-sizes and using different sized learning sets (e.g. 70%, 80% - the remaining 30% and 20% respectively will be the test set).

   To get the multi-variate class conditional probability density function use the the `KernelDensity` class from the Python library `sklearn.neighbours`. You will also need the `GridSearchCV` function from `sklearn.model_selection` to get the best density estimator. This is a non-parametric way to get a density estimator.

   To understand how to do density estimation using the `KernelDensity` class read the `scikit-learn` documentation. Also, there are several tutorials on the internet that should help.

   To think about: What is the complexity of building the BDR classifier? What will happen if the number of attributes and the number of feature vectors in the learning set are large?

   [100]