Project Description

CSE 802 Spring 2020

Dataset

- Select a dataset that has many samples
- Important to consider the number of classes (c) and the number of features (d) in the dataset
- Ideally, you want to select a dataset which has at least 4 classes and 50 dimensions
    - This is an arbitrary number – the point is to work on a “realistic” problem rather than a “toy” problem
- Please review D2L for dataset suggestions

Training Set, Validation Set, Test Set

- Partition your dataset into training set, validation set and test set
- Use the training set to develop the classifier and the validation set to select suitable parameters for the classifier (fine-tuning parameters)
- Use the test set to evaluate the performance of the tuned classifier
- Present your results using a confusion matrix besides reporting the overall classification accuracy
  - Which classes are often confused?
  - What is the variance in classification accuracy when the partitioning exercise is repeated multiple times?
  - What is the impact of imbalanced classes on classification accuracy?
  - ….

Pre-processing the Dataset

- Data normalization:
  - Experiment with different feature normalization schemes
  - For example, should features be normalized in the range [0, 1]
  - Or, should they be normalized using the z-score technique
  - Is normalization needed at all?
- Dimensionality reduction:
  - Experiment with feature extraction (projection) and feature selection (e.g., SFFS) techniques
- You can use already available software to test these options

Experimenting with Different Classifiers

- Test the performance of different classifiers on the dataset
    - See D2L for software packages that can be used
- Which classifier results in the best accuracy?
- Write your own code to design a Bayesian Classifier based on the maximum *a posteri*ori principle
    - Estimate class-conditional PDFs assuming a Multi-Variate Gaussian density function (or any other multi-variate density function)
    - Estimate class-conditional PDFs assuming that features are independent and that every feature variable can be modeled using a Gaussian (or any other function)
    - Use non-parametric density estimation schemes to determine the class-conditional density values of a test sample

The Goal

- Gain insight into the dataset you selected
- Test at least
  - 2 dimensionality reduction schemes
  - 3 classifiers from existing software packages
  - Bayesian classifiers that you'd designed in homework #2, #3, #4
- Test the impact of changing the training data – do this multiple times in order to obtain the mean and variance of the error rate
- Draw some conclusions about the dataset, the features used, stability of various classifiers, etc.
- Submit the report by May 1, 2020
  - Grade will be based on the degree of analysis conducted