CS697A – Topic in Computer Science – Machine Learning
Summer 2020

Assignment 2 (7.5 points)
Due date : July 12, 2020 Sunday at 11:00pm

PURPOSE:

Review: Chapter 3 (Bayesian Decision Theory), Chapter 4 (Parametric Methods), Chapter 5 (Multivariate Methods). Deciding on the right model complexity to prevent overfitting.

WHERE TO SUBMIT ASSIGNMENTS:

Please submit through the class Blackboard site. Please zip and upload all your files using filename studentID_HW2.zip. Submit a zip file of the Jupyter Python notebook you used for your homework.

POLICY:

Collaboration in the form of discussions is acceptable, but you should write your own answer/code by yourself. Cheating is highly discouraged for it could mean a zero or negative grade from the homework. If a question is not clear, please let me know (via email, during office hour or in class). Do not use a library unless it is a very basic one or it is indicated otherwise.

DATA:
Read:
https://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits

Download the dataset and read the description carefully:
https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/
test set: optdigits.tes
training set: optdigits.tra

Use only the data for classes 6 and 9 from the test set. For training you are provided with the files `optdigits6_100.tra` and `optdigits9_100.tra` which contain 100 instances from class6 and class9. Repeat all the experiments below using the first 50 instances and then the whole 100 instances from each class for training.

Questions:

Q1 [2pts]: Parametric Classification: Using each of the 64 input features separately as the single input dimension, use parametric classification, assuming that the input is distributed according to a Gaussian. Report the training and test ~~confusion matrices and~~ errors for the case of each of the 64 features. Which feature(s) give the best test performance?

Q2 [2.5pts]: Use all the 64 features, assume that inputs are 64 dimensional Gaussians, and assume that for each class the covariance matrix is different. Report the training and test confusion matrices and errors. **Hint:** eliminate features that have covariance zero.

Q3 [1.5pts]: Repeat Q2, assuming that all the class covariance matrices are the same.

Q4 [1.5pts]: Use the first 10 features in Q1 that gave the best test performance and repeat Q2. Compare the test performance you got to Q2.