**CS697A – Topic in Computer Science – Machine Learning**

**Summer 2021**

**Assignment 2 (10 points)**

Due date : July 20, 2021 Tuesday 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:**

Fill in expected row and submit as a separate excel file, studentID\_HW2\_expected.xlsx in your zipped file.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Question | Q1 | Q2 | Q3 | Q4 | **TOTAL** |
| Max | 3 | 3 | 3 | 1 | **10** |
| Expected |  |  |  |  |  |

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, the datafiles, expected grade excel file and also a pdf answering the questions for the homework.

Include your name, ID, your groupID and your groupmates’ names and IDs in all files. Each group member must submit the same file.

**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>

For this hw, you are provided with subset of the training and test datasets ((optdigits.tra, optdigits.tes) from

<https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/>

The training and test datasets that you are going to use, optdigits\_69\_N200.tra and optdigits\_69.tes contain only the instances from classes for digits 6 and 9.

**Questions:**

**Hint1:** eliminate (i.e. delete) features that have variance zero for any of the classes in the training set. You should eliminate the columns both from the training set and the test set.

**Hint2:** use the training set to estimate the mean and covariance (matrix) values that you will use to compute your discriminant function.

**Q1 [3pts]:** 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 errors for the case of each of the 64 features. Which feature(s) give the best test performance?

**Q2 [3pts]:** 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.

**Q3 [3pts]:** Repeat Q2, assuming that all the class covariance matrices are the same.

**Q4 [1pts]:** Use the first 10 features in Q1 that gave the best training performance and repeat Q2 using those 10 dimensions (columns). Compare the test error performance you got to Q2.