Statistical Machine Learning

Programming Assignment

Prabhat Kumar, MT17036

Question 6

Dataset

The dataset was found to contain 715 images of file extension (.pgm, ,pgm.bad) of image resolution

Methodology

- Images were read in form of Numpy array using Skimage library.
- Images are resized to a resolution of 50x50, to avoid memory error.
- Image vectors are flattened to get feature vector of size 1x2500
- PCA is applied, by calculating Eigen Vectors of Covariance Matrix, with each vector of size 1x2500.
- Eigen Vectors are extracted depending upon the Eigen Energy and are Matrix Multipled to data to get Transformed_Data.

Observations and Inferences

- As the Eigen Energy is increased no of features of projection data point increases, 14 for 90%, 35 for 95% and 135 for 99%.
- Each eigen vectors highlights some features of faces and eigen values are the measure of importance of the feature highlighted.
- Accuracy of Projected data increases upon increase in Eigen Energy as, more the eigen energy, the less the loss of data from the original data.

Eigen Energy (%)	Accuracy (%)
90	76.53631284916201
95	92.17877094972067
99	95.2513966480447

- Accuracy of original data should be theatrically higher than of those with projected data points as stated above that loss of data is higher with lower eigen energy.

Accuracy on Original Data: 95.53072625698324%

NOTE:

- **Eigen Value Representation as Images can be found in Q6/Output_Eigen_Vect**
- **Projection Matrix of various Eigen Energy can be found at Q6/data with name Projection_<EigenEnergy>.sav**
- **Transformed Representation of data depending upon various Eigen Energy can be found at Q6/transformed_x_<EigenEnergy>.sav**

Question 5

Dataset:

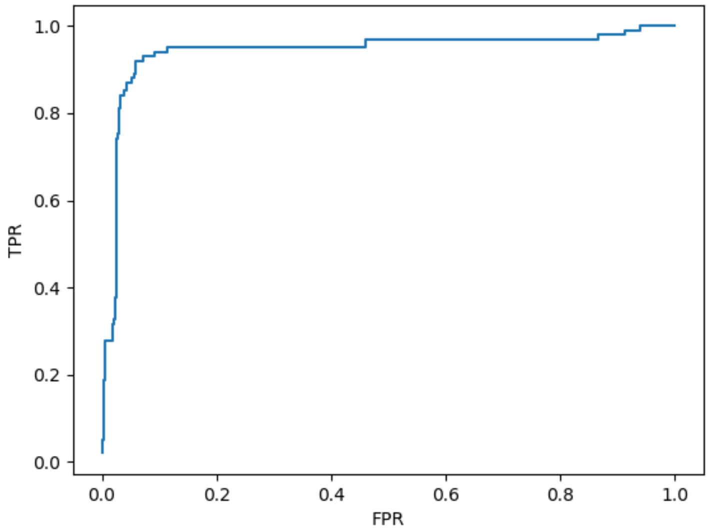
MNIST Dataset contains 60000 images in Training Set and 10000 images in the Test Set. Each image is of resolution 28x28.

Methodology

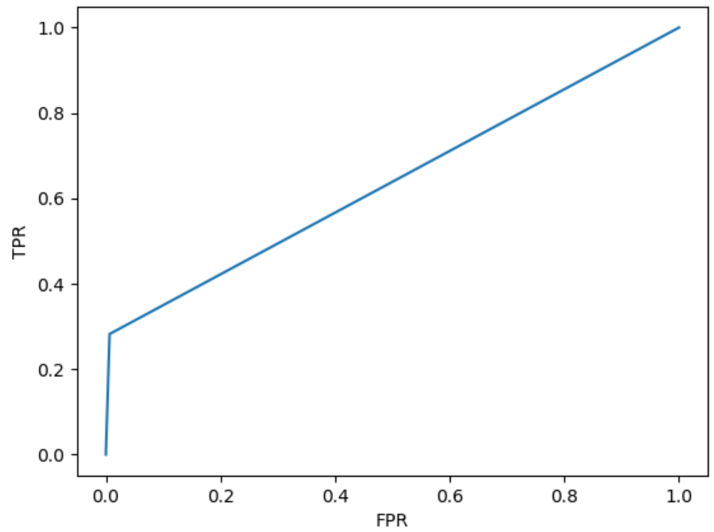
- Each data point of dataset was flattened to vector of size 1x784
- Standard Deviation and Mean is calculated for each of the 784 features per Class.
- Naïve Bayes is implemented by assuming each feature independent of each other.
- Posterior is calculated by assuming each feature is distributed by Gaussian Distribution.

Observation and Inferences

- Data and its features have been assumed to follow Gaussian Distribution because, as stated by Central Limit Theorem, for a large aggregate effect of large number of small, independent random variables leads to a gaussian distribution



ROC Curve for 0-1 Classification



ROC Curve for 3-8 Classification for 10-90 Split

Classification Class	**Split**	**Accuracy**
0,1	No Split	98.58156028368794
3,8	No Split	65.625
3,8	10-90	30.25

- A classifier is said to be unbiased if prior of data on which it is trained upon depicts the real occurrence of data. Here the data is trained upon 10-90 split of occurrence of digit 3 and 8, and so is the test data. Thus, making the classifier unbiased against the test set. Whereas compared to the whole dataset the classifier is biased.

Question 2.8-7

- True Error has been calculated in terms of erf(.) function by transforming P(error) in terms of Cumulative Distribution Function of Normal Distribution Probability Density Function

$$\text{CDF(x)} = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-u}{\sigma\sqrt{2}}\right)\right)$$

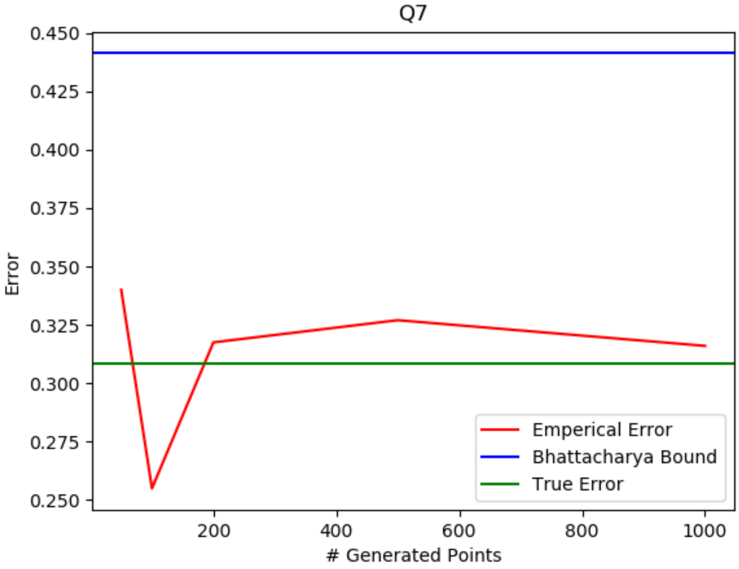
where u is the mean and σ is the standard deviation of the Normal Distribution.

$$\text{CDF(x)}=\int_{-\infty}^{x} p(x)dx$$

- Empirical Error has been calculated by dividing points misclassified by total points.

Empirical Error = # Misclassified Points/# Total Points

Error Type	**Error Value (0,1)**
Bhattacharya Bound	0.4412484512922977
True Error	0.308537538725987

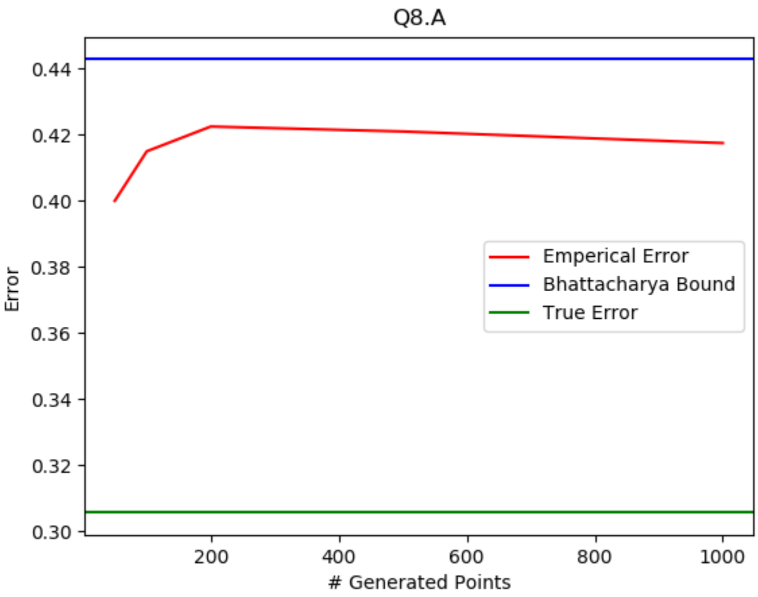


- Empirical Error is found to revolve around the True Error value whereas Error by Bhattacharya Bound is found to be upper bound to both Empirical and True Error

Question 8

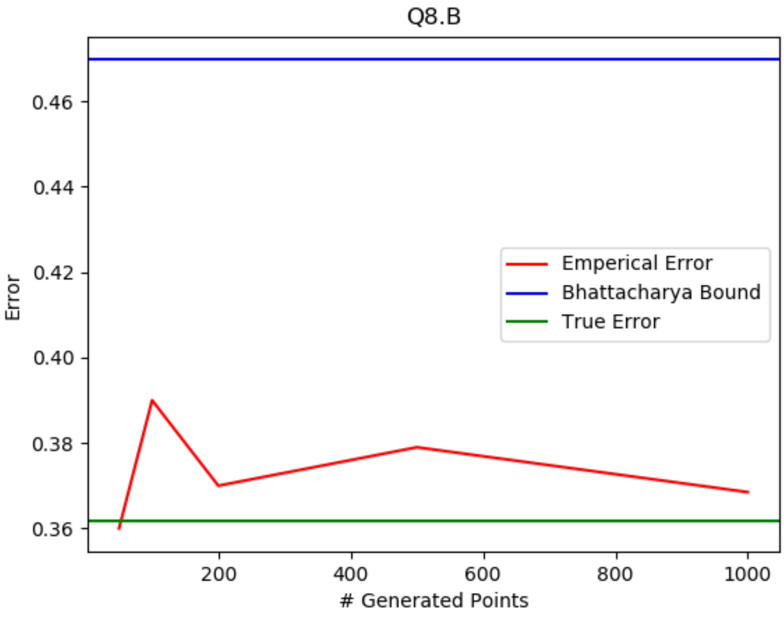
A.

Error Type	Error Value (0,1)
Bhattacharya Bound	0.4428435647004219
True Error	0.3056130786969188



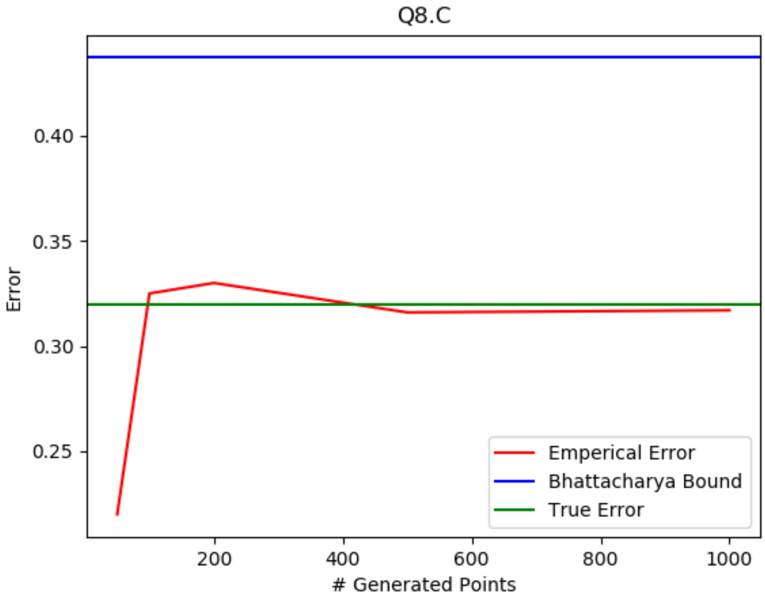
B.

Error Type	Error Value (0,1)
Bhattacharya Bound	0.4697065314067379
True Error	0.36183680491588155



C.

Error Type	Error Value (0,1)
Bhattacharya Bound	0.4371111804791032
True Error	0.3197148574567994



Question 2

- 30 data points has been considered as per Exercise, along with the label W1,W2,W3 which has been kept in file 'book_data.txt'
- If Prior of a class is given as 0, those data points have not been considered for further calculation and estimation.
- Empirical Error has been calculated as per formula given in Q7.
- Bhattacharya Bound has been calculated as per Equation () as per reference [1].

Feature Considered (i at X_i)	**Bhattacharya Bound**	**Empirical Error**
1	0.47399943544659356	0.30
1,2	0.46046616491760517	0.45
1,2,3	0.4119256313835016	0.15

- Empirical Error does not show any trend upon number of Features considered, where Bhattacharya Bound decreases upon increase in number of feature considered.
- Yes, it is possible that empirical error is greater than the Bhattacharya Bound as empirical error is an experimental error upon misclassification of data, as empirical error is purely dependent upon the type of data provided and it may be possible that the given data cannot be mapped by a Normal Distribution, whereas Bhattacharya Bound is a theoretical error which is constant of given set of prior, mean and covariance matrix.

Question 4

Mahalanobis Distance

Data Point	**Class 1**	**Class 2**	**Class 3**
[1,2,1]	1.01497	0.8580	2.6747
[5,3,2]	1.5571	1.7556	0.6470
[0,0,0]	0.4899	0.2684	2.2415
[1,0,0]	0.4872	0.4518	1.4623

Data Points	**Prior = [0.33,0.33,0.33]**	**Prior = [0.8,0.1,0.1]**
[1,2,1]	2	1
[5,3,2]	3	1
[0,0,0]	1	1
[1,0,0]	1	1

References

- Sklearn Documentation : http://scikit-learn.org/stable/documentation.html
- Numpy Documentation : https://docs.scipy.org/doc/
- Richard O. Duda, Peter E. Hart and David G. Stork , Pattern Classification, 2nd Edition