Assignment 1
ELL409 : Machine Intelligence and Learning

Siddharth Khera
Indian Institute of Technology, Delhi
2015MT60567
mt6150567@iitd.ac.in

Utsav Sen
Indian Institute of Technology, Delhi
2015MT60569
mt6150569@iitd.ac.in

Navreet Kaur
Indian Institute of Technology, Delhi
2015TT10917
tt1150917@iitd.ac.in

I. Introduction

We have used different classification schemes like Bayes, Naive Baye, K-means and K-NN to perform classification on three different datasets. For the Bayes classifier, parametric technique of Maximum Likelihood Estimation and the non-Parametric technique of K nearest approximation was used for predicting the class Conditional Densities(CCDs). Principal Component Analysis(PCA) was applied, with and without whitening, for dimensionality reduction on datasets with large dimensions. For parametric estimation, form of CCD was assumed by visualising the distribution of features for each class. As K-means does not guarantee global optimum and the resulting clusters are dependent on the initialisation, a number of random initialisations were done and the one with highest accuracy was chosen. We also experimented with different distance measures like Euclidean, Manhattan, Chebyshev and Mahalanobis for K-Means and K-NN clustering.

II. Classification

A. Fashion MNIST Dataset

Fashion-MNIST dataset consists of 60,000 training examples and 10,000 test examples. Each example is a 28x28 pixels gray-scale image. Each image is labeled with 10 class categories. Each image is considered to be 784 dimensional data sample. Figure 1 shows the distribution of data projected on a 3D space.

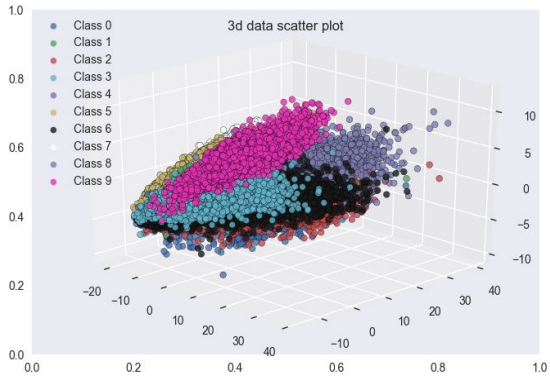


Figure 1: Distribution of data in 3-D space

As part of preprocessing the data set, PCA was done to reduce the dimension of the data. Figure 2 shows the variation of variance retained and accuracy of classification model with the number of components. It is noticed that the growth in accuracy for Naive Bayes classifier becomes stagnant beyond 80 principal components, at which maximum accuracy is achieved. Most of the variance in data is captured by these 80 components and inclusion of more features does not add to accuracy of the model.

Correlation heatmaps of these 80 components and distribution of most important feature for each class are shown in Figure 1 and 2 in Appendix. As clear from Figure 1, some of the features are fairly correlated for classes 1, 3, 7 and 9. It can be said that if were to use the Bayes classifier without the Naive assumption, results for these classes could have been better.

PCA was applied with and without whitening. However, no significant difference between the accuracies for the two cases was observed.

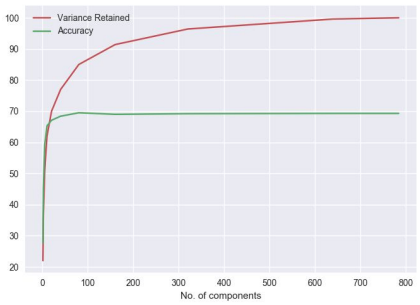


Figure 2: Accuracy of classification and Variance Retained v/s Number of Principal Components

Different classification schemes used for this task and their accuracies are shown in Table 1. These accuracies correspond to the highest accuracies obtained by varying number of components and trying out different CCDs and estimation techniques for Bayes and Naive Bayes (Multivariate Gaussian distribution, MLE, K-NN estimation), varying distance metrics(Manhattan, Euclidean, Chebyshev, Mahalanobis) for K-means and K-NN and varying *k* for K-NN.

Model	Parameters	Train Accuracy	Test Accuracy
Bayesian	#PC* = 80	71.89	70.51
Naive Bayes	#PC = 80	71.26	69.51
K - Means	PCs = 10, n** = 2	48.57	46.90
K- NN			

Table 1. Classification Schemes and their Accuracies

* #PC = no. of principal components

** n-number of initialisations

Maximum accuracy is obtained by the Bayes classifier, with Naive Bayes' accuracy being slightly lower than it.

Bayes is not preferred since computing the multivariate distribution for 784 features itself is a very costly operation. The time taken by it to predict the classes is outweighed by the accuracy it achieves as compared to Naive Bayes with PCA.

Note: On using Sklearn PCA, accuracy of this naive bayes shot up to 80.

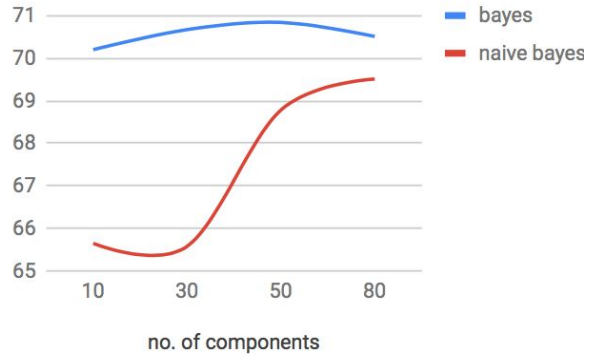


Figure 3: Performance of Bayes and Naive Bayes with number of principal component

Since the number of dimensions is fairly large, our accuracy for K-means and K-NN is bounded by the 'curse of dimensionality' due to which it does not give good results as the distance measures do not make sense in high dimensions because when the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse.

The confusion matrix for training and test data using Naive Bayes is shown in Figure 4(a), (b). It can be inferred that all the classes are fairly separable except coats and shirts(class 4 and 6), pullovers and coats(class 2 and 4), and sandals and sneakers(class 5 and 7) as the values corresponding to these points have a high value(>200) in the confusion matrix. It can be noted that $C(i,j) \neq C(j,i)$ i.e, the classifier predicts class 6 as 4 more often than vice versa.

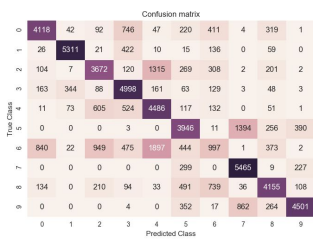


Figure 4(a): Confusion matrix on training data

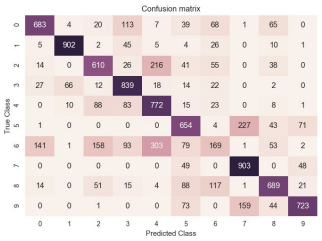


Figure 4(b): Confusion matrix on test data

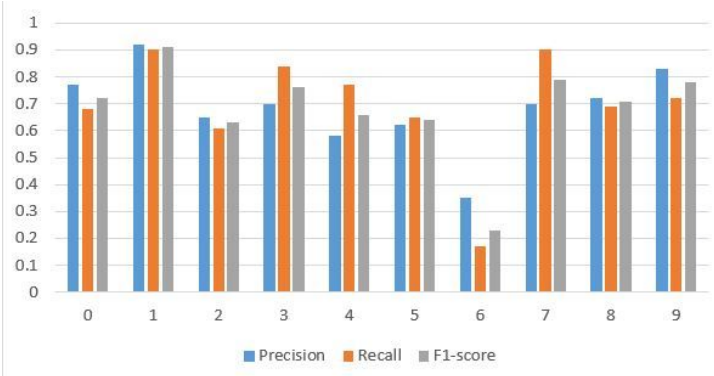


Figure 5: Performance Parameters

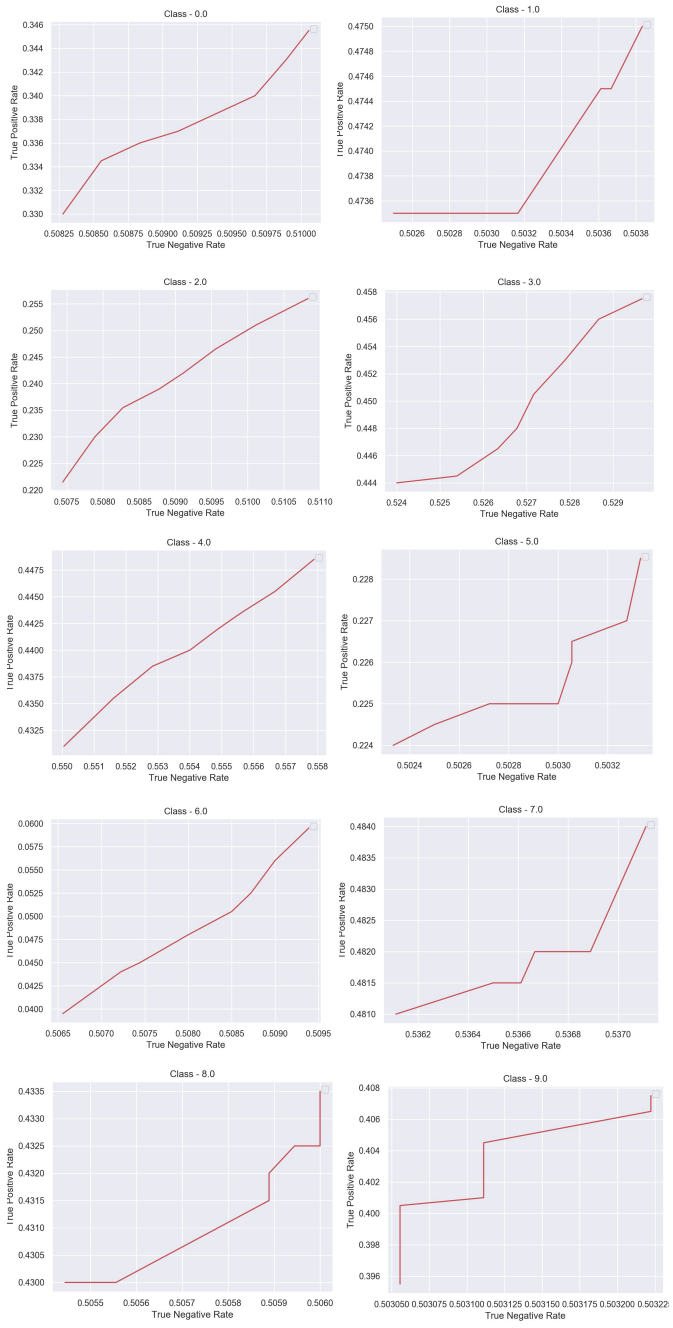


Figure 6: ROC Curves

B. Blood Test Dataset

This dataset consist of outcomes of three Blood Tests (Test1, Test2 and Test3) for analyzing the condition of Heart of a patient. It contains doctor's advice on whether the heart is HEALTHY, MEDICATION (patient needs medication) and SURGERY (if there is a need of any kind of Surgery) based on the outcomes of the three tests.

It can be seen from Figure 7 that the classes are not well separated and the shapes of the clusters are not spherical. In fact, cluster of class 1(Medication) is of an elliptical and flat shape and has a large radius. Class 0(Health) looks well separated from 1 and 2(Surgery) but clusters of 1 and 2 are slightly intermingled.

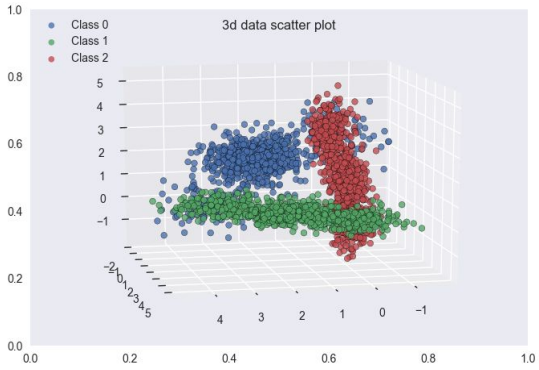


Figure 7: Distribution of data in 3-D space

A consequence of such a data distribution can also be seen in the results of K-means classifier which achieves an accuracy of about 60% only . The correlation heatmap of features for each class is shown in Figure 8. The features are fairly uncorrelated, the consequence of it being that there is no significant difference between the performance of Bayes and Naive Bayes.

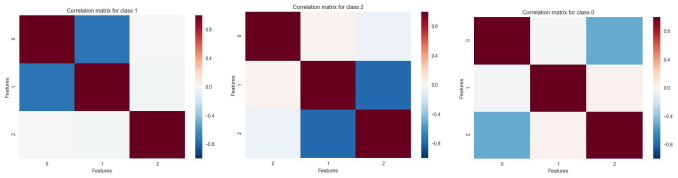


Figure 8: Correlation Heatmap on Test data

Accuracies obtained from different classification schemes for this dataset are shown Table 2. Bayes and Naive Bayes accuracies correspond to CCDs of Multivariate Gaussian and MLE. While K-means corresponds to $k = 3$ and Manhattan distance, $k = 9$ K-NN which is evaluated on Euclidean distance.

Model	Train Accuracy	Test Accuracy
Bayesian	90.47	89.81
Naive Bayes	90.66	89.86
K - Means	77.84	77.1
K- NN*	89.90	89.433

Table 2. Classification Schemes and their Accuracies

*with k = 9

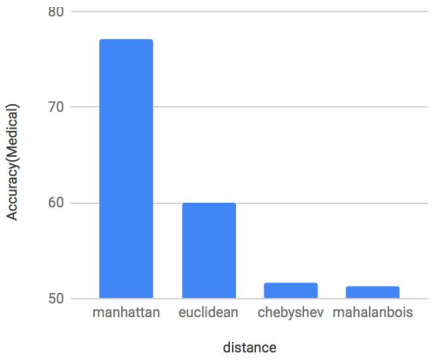


Figure 9: Performance of K-Means for different distance metrics

Different distance metrics were used for k-means clustering. It was found that maximum accuracy of 77.1 was achieved on training data with Manhattan distance.

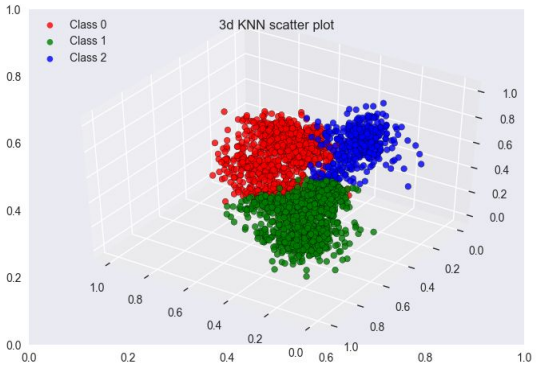


Figure 10: Clusters obtained from K-means

Red	Class 0 : 706 ; Class 1 :306	Assigned 0
Green	Class 1 : 505; Class 2 : 474	Assigned 1
Blue	Class 0 : 103; Class 2 : 301	Assigned 2

Table 3: True Distribution of Clusters

K-means is incapable of predicting clusters that are not spherically shaped, as in this case. It can be noticed from Figure 10(a) that the predicted clusters are spherical in shape, however, the true clusters are of different shapes, with class 1 being an elliptical flat cluster as noticed in Figure 7. True distribution of clusters is shown in Table 3.

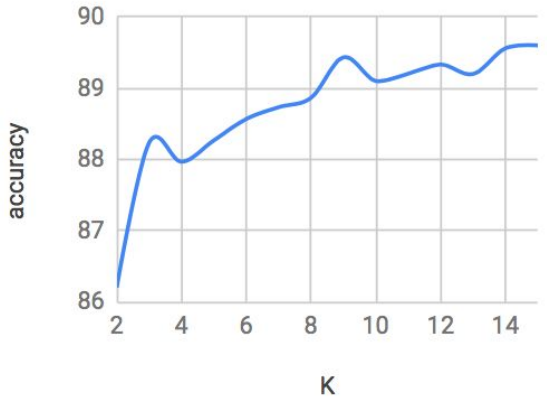


Figure 11. Accuracy v/s no. of nearest neighbours(K)

Deriving inference from Figure 8, the value of the parameters k denoting number of nearest neighbours was chosen to be 9. The reason being that 9 is located at a local maxima, hence appears to be the natural neighbourhood of the data. On further increment to k, we over-smoothen the data.

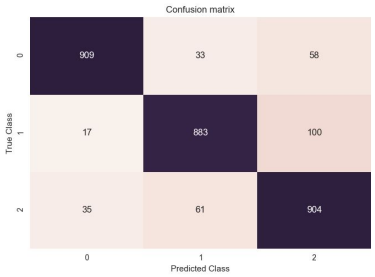


Figure 12. Confusion matrix on test data

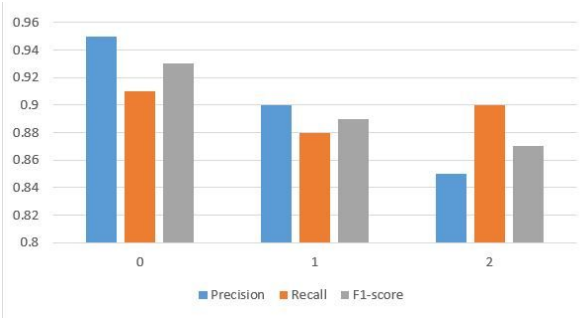
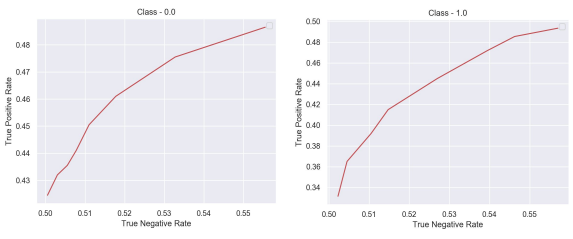


Figure 13. Performance Parameters



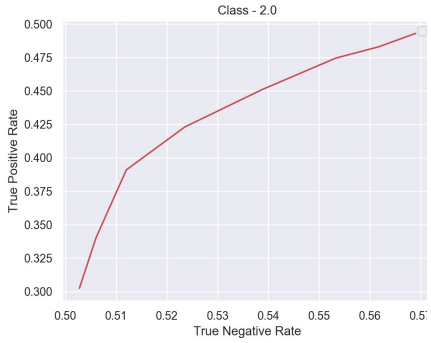


Figure 14. ROC curves

B. Train Selection Dataset

The dataset contains all the information about the person booking the train along with whether the person has boarded the train or not. Classifiers are built to predict whether a person will board the train or not if provided with information such as age, fare paid, number of members traveling with etc.
Accuracies obtained from different classification schemes for this dataset are shown Table 4. While K-means corresponds to $k = 2$ and Manhattan distance, $k = 15$ is used for K-NN which is evaluated on Mahalanobis distance. This is reasonable since the data mostly consists of discrete features for which Mahalanobis is a good distance metric as compared to others.

Model	Train Accuracy	Test Accuracy
Bayesian	78.51	77.10
Naive Bayes*	75.95	75.54
K - Means	64.46	64.00
K- NN**	80.60	80.1

Table 4. Classification Schemes and their Accuracies

* Assuming Gaussian CCDs on continuous variables and Multinomial on categorical variables {budget: Gaussian, number_count: Gaussian, sex: Multinomial, preffered_class: Multinomial, Age: Gaussian}
** with k = 15

A peculiar observation for this dataset was that with the same test and train split ratio, the variance in the accuracy was high subject to uniform sampling of test data. This suggets lack of appropriate features in the dataset as well as the small size of training set.

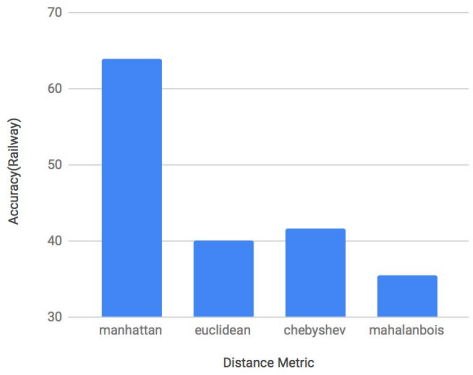


Figure 16: Performance of K-Means for different distance metrics

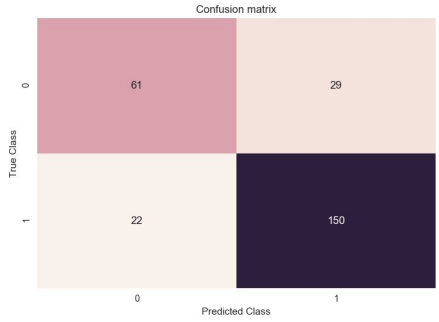


Figure 15. Confusion matrix on test data

From Figure 15, it can be observed that prediction for both classes was fairly accurate. Also, one class in the data (boarded) clearly dominates in frequency the other class (not boarded).

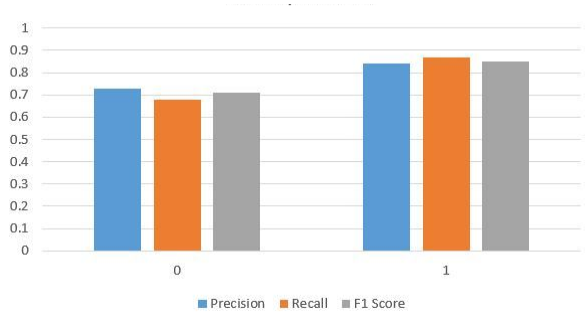


Figure 17: Performance Parameters

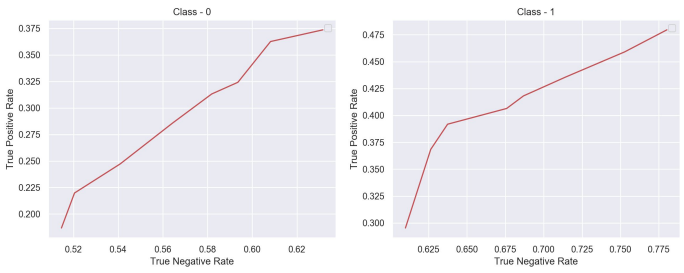


Figure 18: ROC Curves

Appendix

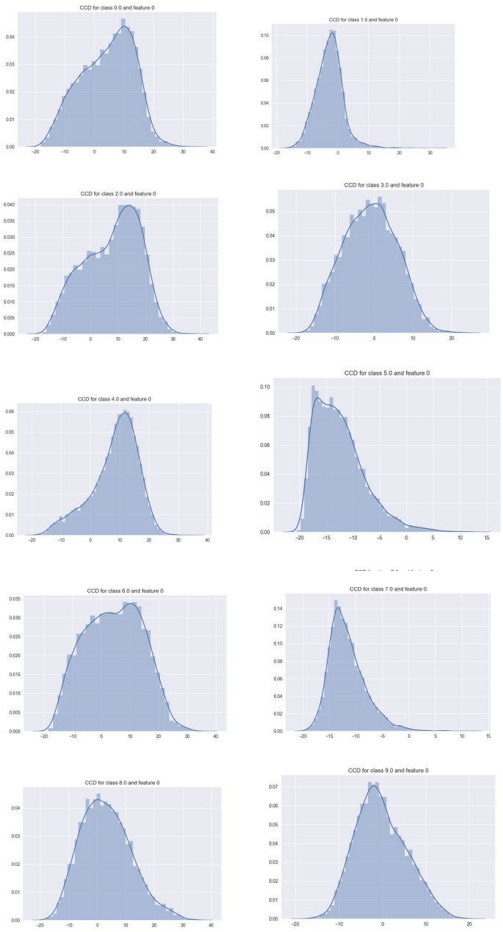


Figure 1: Distribution of most important feature for all classes (F-MNIST dataset)

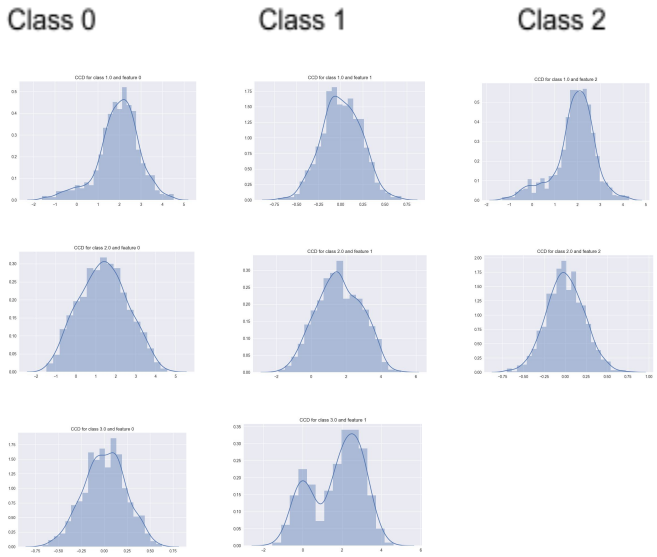


Figure 2: Distributions of the features (Medical Dataset)

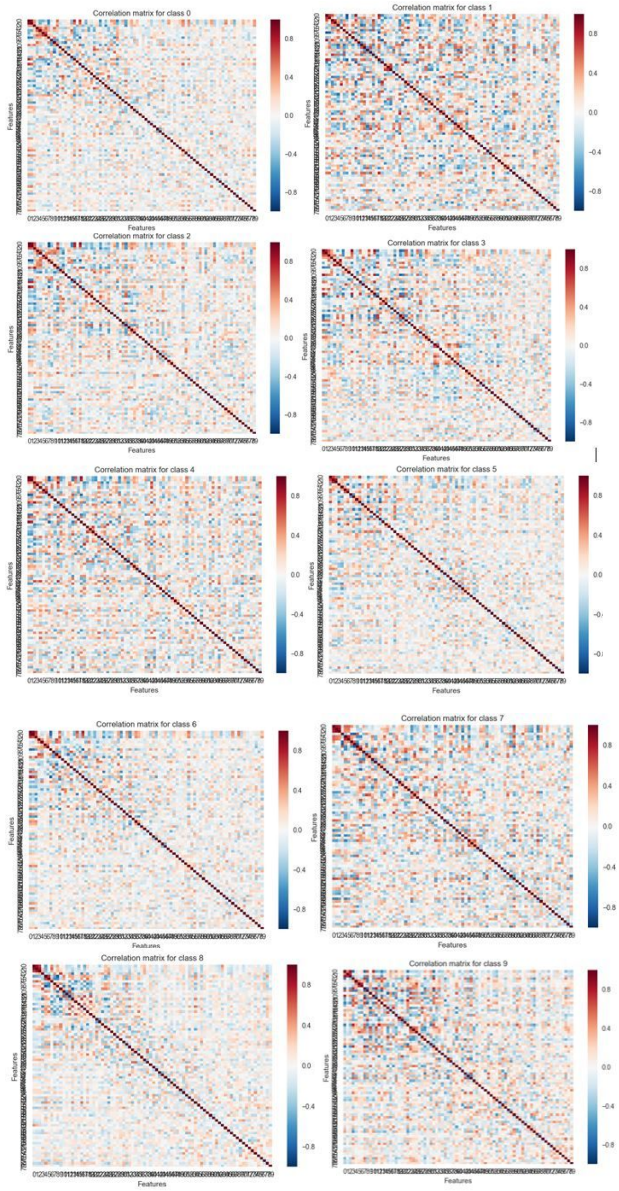


Figure 3: Correlation Heatmap of Features (F-MNIST dataset)

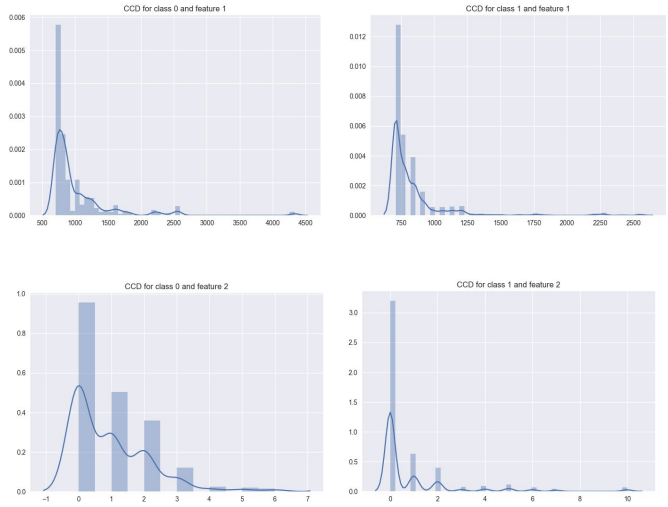


Figure 4: Distributions of the features (Railway Dataset)