Assignment 1

ELL409 : Machine Intelligence and Learning

Semester-1 2018-19

Instructor - Prathosh AP

1 Problem statement

The *Assignment 1* consists of the tasks of building up different kinds of classifiers on various example datasets. Information about these datasets is provided in forthcoming sections. It is required to implement following types of Classification schemes for each problem and record your observations.

1. Bayes Classifier (with different class conditional densities and estimation techniques)
2. Naive Bayes Classifier
3. K-means Clustering
4. K-Nearest Neighbor Classifier
5. Principal component analysis (where ever applicable)

Use *Python* as your programming language for implementing the code. It is highly expected not to use the available APIs for getting the results but to implement the classifiers with you own code. You can always verify your results with the respect to the outcome of available APIs but you should not be using or copying the code from various resources available online or elsewhere. Please be aware that these exercises are for your own practical understanding of various classification techniques. By merely using APIs or copying code will not benefit you in long run.

It is highly advised to implement the code with standard design guidelines (e.g. modular, object-oriented, re-usability etc). This will help in extending your code for more complex tasks and also while testing or debugging the code. The quality of the code will also be given weightage in overall evaluation of the assignment.

You need to implement your test code for each of the classifier. Also, compute various Classification Performance Parameters such as *Accuracy*, *Recall*, *Precision*, *F1-Score*, *ROC-curve*. (*design Hint: If you can implement these functions under a class or as various functions, then these can easily be used for all kinds of classifiers*).

2 Expected Outcome

For each of the datasets, you should compare the classification performance for different classification methods.

You need to provide the analysis report for the results observed and various conclusions. Report should be in standard IEEE format not exceeding 4 double column format. Assignment will be graded in accordance with the quality of the analysis that you carried out and the quality of the report. There is no upper limit to the number of experiments that you can carry out (such as using different forms for class conditional densities, estimation techniques, if you are using GMM/K-means, number of mixture components, different initialisation techniques etc.) - It is only limited by ones imaginative and intuitive power. But the minimum expectation is to implement all the algorithms from scratch under different parameter settings and record the outcomes.

You are supposed to submit the code too - I suggest you to submit the Jupiter notebook file so that evaluation will be easier.

3 Datasets

3.1 Fashion-MNIST Dataset

Fashion-MNIST is a dataset consisting 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. Figure 1 shows example images in this dataset.

Here, each image is considered to be 784 dimensional data sample. So, there is a need to reduce the dimension of the data. Principal component analysis can be used for selecting the important features and create a lower dimensional feature vector for classification task. So, implement a module or a function to get the principal components for each of the data sample. Again, implement your own code for implementing PCA. Also, evaluate the variation of classification performance with variation in number of principal components that are included as part of feature vectors.

You can retrieve this dataset from `https://github.com/zalandoresearch/fashion-mnist`

3.2 Blood Test

This dataset consist of outcomes of three Blood Tests (Test1, Test2 and Test3) for analyzing the condition of Heart of a patient. Doctors in a hospital is analyzing these outcomes and are providing the report for patient indicating whether the Heart is Healthy or the patient needs medication or there is a need of any kind of Surgery. This dataset is also containing the doctor's advise for whether the Heart is *HEALTHY*, *MEDICATION* and *SURGERY* based on the outcomes of the three tests.

You are required to create various kinds classifiers classifying patients in categories of *Healthy*, *need medication* or *undergo surgery*.

The snapshot of the dataset is shown in Figure 2 and the same will be made available to you with this document.

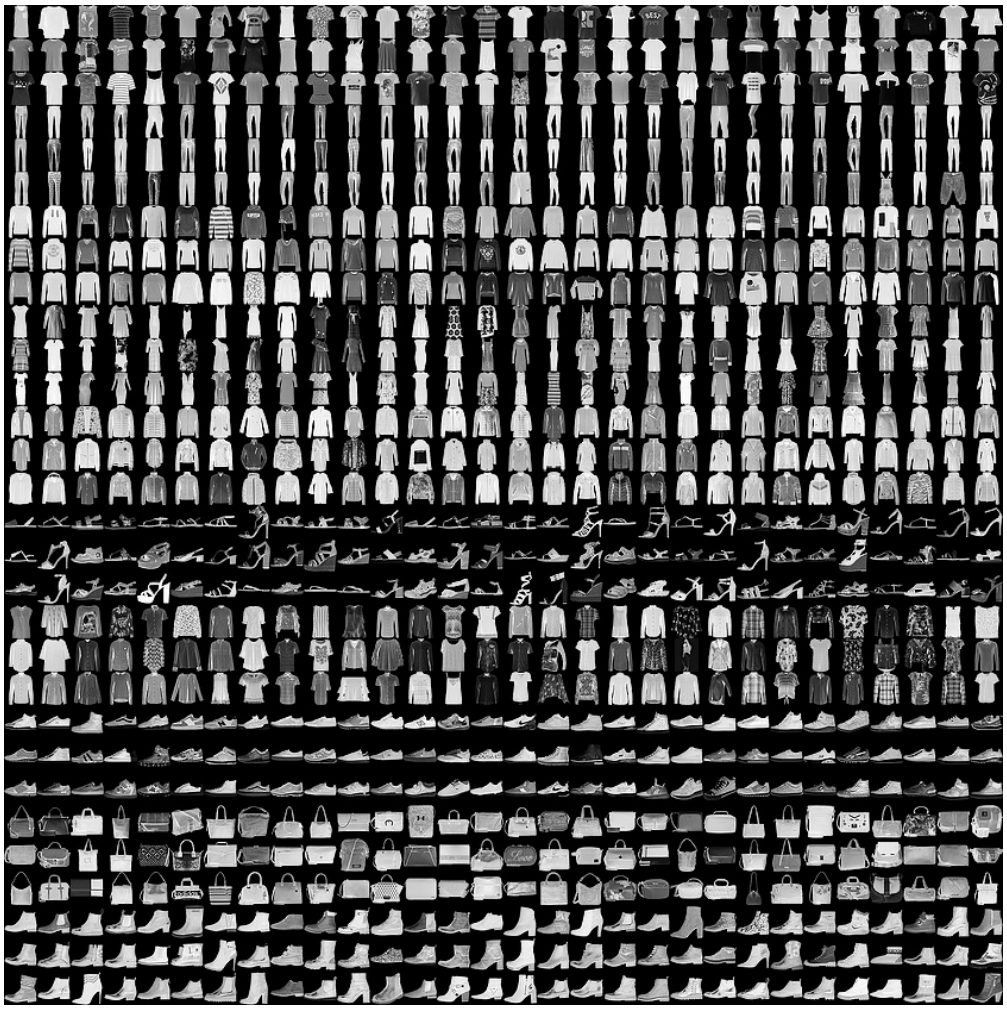

Figure 1: Fashion-MNIST Dataset

Health	TEST1	TEST2	TEST3
MEDICATION	4.309531	-0.83201	0.051151
HEALTHY	2.452432	0.067027	1.825669
SURGERY	-0.16333	1.827832	-0.38598
HEALTHY	2.069746	-0.0386	2.77623
SURGERY	-0.32149	1.982645	1.475755
MEDICATION	1.289905	2.202437	0.129948
HEALTHY	1.875493	-0.22852	2.033736

Figure 2: Heart Health Test Dataset

3.3 Train Selection

Indian Railways has introduced a new luxury train from Mumbai to New Delhi. This train has all facilities like WiFi, Club, Lounge, Playing, SPA etc. Each of the facilities are chargeable along

caseID	Whether boarded the Train?	Train Fare to be Paid	Number of Family Members travelling with	preferred Class	sex	Age Category
111131089	0	2201	0	FIRST_AC	female	2
2489059216	0	1775	3	FIRST_AC	male	0
1565109576	1	1775	3	FIRST_AC	female	0
1373075087	1	1775	3	FIRST_AC	male	3
1598041082	1	1775	3	FIRST_AC	female	2
3825576434	0	852	0	FIRST_AC	male	4
644658844	0	1207	1	FIRST_AC	female	6
688816386	1	710	0	FIRST_AC	male	3

Figure 3: Train Selection Dataset

with the travel fare. To analyze the interest shown by public, they floated a form with information such as Age, Sex, fare paid, number of members traveling with, Travel class etc. This form was filled by the person while booking the ticket for the train. After the first day launch of the train, the department analyzed whether the person has boarded the train or not.

The dataset that is provided contains all the information about the person along with whether the person as boarded the train or not. You need to create classifiers to classifying whether a person will board the train or not if provided with information such as age, fare paid, number of members traveling with etc.

The snapshot of the dataset is shown in Figure 3 and the same will be made available to you with this document.