**Project – 4**

**Problem 1: PCA and Feature Selection**

**1A: SVMs and PCA**

1. Top 6 Eigen Values of the data covariance matrix is given below.

[10362.43060735, 8000.98321771, 6851.81986497, 4939.33641135,

4117.61132227, 3699.1664645]

1. **Train Error Matrix K vs C:**

[[40.8974359 40.8974359 40.8974359 40.8974359]

[42.37179487 42.37179487 42.37179487 42.37179487]

[40.57692308 40.57692308 40.57692308 40.57692308]

[38.78205128 38.78205128 38.78205128 38.78205128]

[38.71794872 38.65384615 38.65384615 38.65384615]

[38.46153846 38.46153846 38.46153846 38.46153846]]

**Validation Error Matrix K vs C:**

[[40.8974359 40.8974359 40.8974359 40.8974359]

[41.02564103 41.02564103 41.02564103 41.02564103]

[43.07692308 43.07692308 43.07692308 43.07692308]

[39.87179487 39.87179487 39.87179487 39.87179487]

[38.84615385 39.1025641 39.1025641 39.1025641]

[39.35897436 39.35897436 39.35897436 39.35897436]]

**Test Error Matrix K vs C:**

[[41.15384615 41.15384615 41.15384615 41.15384615]

[43.84615385 43.84615385 43.84615385 43.84615385]

[43.07692308 43.07692308 43.07692308 43.07692308]

[38.07692308 38.07692308 38.07692308 38.07692308]

[41.15384615 41.15384615 41.15384615 41.15384615]

[40.76923077 40.76923077 40.76923077 40.76923077]]

1. The best (K, C) pairs according to the validation accuracies are K = 5 and C = 1.
2. The best (K, C) pairs according to the test accuracies are K = 4 and C = 1.

**1B: PCA for Feature Selection:**

1. Top 11 Eigen values are:

[10362.43060735, 8000.98321771, 6851.81986497, 4939.33641135,

4117.61132227, 3699.1664645, 3609.7266692, 3590.43218958,

3578.58400368, 3538.95445784, 3526.49438646]

1. **A picture containing text, document

   Description automatically generated**

Text, letter

Description automatically generated

1. **C)**

The best error on Train, Validation and Test data by performing PCA and selecting features is as   
follows:

As we can see that this is a very reasonable estimate for SVM without feature Selection. The accuracies are comparable.

**Pros for this approach:**

1. We can identify the best attributes of the given data. This helps in understanding the data better.
2. It reduces the dimensionality of the dataset.
3. The approach helps in better visualization of the data.

**Cons for this approach:**

1. There is a possibility of getting very bad accuracies for some columns.
2. Algorithm needs to be run multiple times to get into a conclusion. This is not helpful for   
   computation.
3. Model developed with this approach may be susceptible to outliers.

**Problem 2: Spectral Clustering and KMeans**

**The Basic Algorithm:**

**A picture containing text, document

Description automatically generated**

**Text, letter

Description automatically generated**

1. **Simple Comparison:**

We can see from different runs that the Kmeans algorithm is not able to converge properly.   
whereas the spectral clustering algorithm converges and able to split the concentric circles.  
  
For Sigma = 0.05 Spectral clustering outperforms the normal KMeans clustering.

1. **Partitioning Images:**

Using K means Clustering we are not able to partition the image accurately. The inner loops are not being detected properly.

Whereas with Spectral Clustering, the detection is somewhat better but not accurate. We need more info for this to detect properly.