EEL709/ELL784

*Assignment 3*

*Unsupervised Learning*

Pranit Arora

2012ME20861

I. Introduction

The objective of Unsupervised Learning is to cluster or group the available data based on certain feature similarities or patterns. K-means clustering and Gaussian Mixture Models are two widely used methods for the same. Both these methods help cluster the data, however the GMM is a more generative (probabilistic) approach. Principal Component Analysis (PCA) was used to reduce the dimension of the data prior to clustering. This is effective as it captures the important features and reduces the computational cost. It can be defined in two ways. One is that it is the orthogonal projection of the data onto a lower dimensional subspace (principal subspace), such that the variance of the projected data is maximised. It can also be defined as the linear projection that minimizes average projection cost, defined by the Euclidian metric.

II. K-Means Clustering

K-Means clustering with K = 10 was run on the handwritten digits dataset. It was run 5 times on the dataset, using the Eucleadian metric as the distance measure. The accuracies obtained were 48.30%, 55.40%, 61.20%, 55.35%, 50.50 respectively. From the results, we can see that the clusters are mostly based on the appearance of the digits.

The first cluster has 6.

Second cluster mostly 7 and a few 9’s.

Third cluster has mostly 0s with a few 5’s & 3’s.

Fourth cluster has a majority of 0’s.

Fifth cluster has 4s and some 9’s.

Sixth cluster has 8’s, with a few 5’s and 3’s.

Seventh cluster has mostly 4’s and 9’s, with some 7’s.

Eighth cluster has mostly 1’s.

Ninth cluster has 3’s with some 5’s and 8’s.

10th cluster has a majority of 2’s.

Hence we see maximum misclassification happens where a digit could be represented by potentially many clusters or where no cluster clearly represents a particular digit.

From the data, we see that 1st, 4th and the 10th clusters have very high accuracies of 80.82%, 88.39% and 91.45% respectively. This is because these clusters have grouped most of the 6’s, 0’s and 2’s into it by closely identifying their centre. We can also note that the 8th cluster, though it is able to capture a massive majority of 1’s(98.5%), it still has lower accuracy (58.28%) as it also captures a large number of other numbers in the clusters .This is because it is a loose cluster.

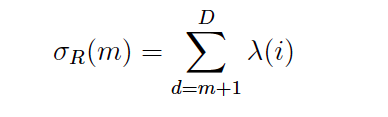
We see that in the case of K=5, more digits are clubbed to one centroids. The similar kinds of digits have been grouped together. For instance, the digits 4’s, 7’s and 9’s have been grouped together. 3’s, 5’s and 8’s have been grouped together. 1’s and 0’s are clearly visible as separate clusters. The remaining digits like 2’s, 6’s and some other digits have been grouped together.

III. Dimensionality Reduction

Principal Component Analysis (PCA) was used for dimensionality reduction prior to K-Means Clustering. The number of principal components were selected such that the residual variance was under 10%.

Residual Variance for the first ’m’ principal components

is defined as :



where λ denotes the possible eigen values and are arranged in descending order.

The minimum number of components that maintained the residual variance under 10% was 83. K-means clustering was further done using 83 principal components. The graph for residual variance can be plotted and be seen to be an exponentially decreases curve, with initial few components each capturing 5-10% of the variance, and the final 700 elements cumulatively capturing only 10% of the variance.

We see that the principal components are capturing different features. Some seem to be capturing rotation and translation, while others are capturing more complicated details.

K-Means clustering was run in PCA space using 83 principal components (10% residual variance criteria). It was observed that the performance was same as that was obtained using 784 raw pixels. However, the fact that instead of 784 features only 83 features were employed, makes this method much faster. This is because in a digit, the possible degrees of freedom are only 3, namely vertical translation, horizontal translation and rotation. This is assuming there is not much difference in the way digits are written. Hence any digit can be adequately represented using 3 dimensions. For 10 digits, this accounts to 30 dimensions. Assuming variance in the writing style, such that on an average each digit could be written in 3 different ways, the total dimensions required to capture all the essential features would be 90. This is the intuitive idea behind Principal Components.

The fact that 83 Principal Components gives results no worse than all the raw pixels, supports this intuition. In fact as the convergence rate is also faster, the chances of achieving a bad clustering and hence getting lower accuracy decreases.

The labeling accuracies achieved in 10 iterations using PCA with 10 clusters were 52.20%, 56.00%, 52.45%, 60.55%, 55.35%, 50.50%, 61.05%, 54.85%, 55.30% and 61.00% respectively. For 5 clusters, again PCA gives results faster than we get without any dimensionality reduction, while giving similar clusters and clustering accuracies.

IV. GMM Clustering

GMM Clustering was also used to cluster the data. The initial clustering was performed where the centroids, co-variances were initialised using the K-Means result and uniform prior probabilities. Using numpy’s GMM clustering resulted in an accuracy of 67.30% which is an improvement over the K-Means accuracy. For the self-implemented GMM, the convergence criteria was kept as when all components of the mean of the clusters changes by less than 0.01.

The implementation was based on the Bayesian approach as mentioned in the book by Bishop. Using this methodology helps avoid the problem where clusters are formed around a single point and hence the Gaussian can have 0 variance if the metric used is the log likelihood of the distribution.

V. Part 2

For the 2nd part of the assignment, the K-means and numpy’s GMM were used. The parameters and the number of clusters were varied for both models. It was observed that GMM in this case gave no additional accuracy than the K-means. The k means gave maximum accuracy for k=20 clusters, with accuracy decreasing with addition and removal in the number of clusters. Due to lack of time, not all parameter combinations of GMM were tried out. It is likely that with appropriate parameter settings, GMM might give higer accuracies than K-means.

VI. Conclusions

From the analysis carried out, we see that K-means and GMMs are effective tools for unsupervised learning by making clusters. Reduction of dimension using PCA reduces the computational cost and eliminates redundant data. If we compare the feature selection the performance of PCA is better. This can be attributed to the fact than when selecting features from the raw features, it is possible that many features have little information, which combine to make up the supervised learning model. Also, there could be many features that are correlated resulting in little addition to the knowledge gained. However, PCA constructs new features in such a way that they are uncorrelated and captures the maximum independent information, ranked in order of the importance of the information obtained by them. Hence, it would be a useful tool to select features in supervised Learning Models.