CS5691: Pattern Recognition and Machine Learning

Assignment 1

Report

Course Instructor: Arun Rajkumar

Submitted By: Vennapareddy Abhigna

Roll Number: ME19B059

Date: 28/02/2022

Indian Institute of Technology Madras

Chennai 600036, India

Questions and Solutions

1. Given a dataset with 1000 data points each in R^2 i.e., {x1,x2,...,xn} for n=1000 are the data points with 2 features; feature1(denoted as 'x' in dataframe in code file) and feature2(denoted as 'y' in dataframe in code file)

I. Process used in performing PCA in code file is as follows:

- Performed data centering by calculating mean of each feature and subtracting it from the data point values. Stored this new centered data in X with each data point as a column vector.
- Obtained covariance matrix 'C' by $\frac{1}{n}(XX^T)$.
- Calculated eigenvalues and eigenvectors of matrix C. They are:
  - λ_1 = 17.131914402444362 with corresponding eigenvector (-0.323516 , -0.9462227) = w1(principal component 1)
  - λ_2= 14.489604749330638 with corresponding eigenvector (-0.9462227, 0.323516) = w2(principal component 2)

  These eigenvectors form the principal components of their corresponding eigenvalues with w1, w2,.. being the eigenvectors of obtained largest eigenvalues in descending order.

- Eigenvalues obtained are the variance of the corresponding principal component.
  - Variance corresponding to the principal component pc1 [-0.323516, -0.9462227] is **17.131914402444362**.
  - Variance corresponding to the principal component pc2 [-0.9462227, 0.323516] is **14.489604749330638**.
- Variance obtained from pc1 accounts for **54.17802452885222%** and that obtained from pc2 accounts for **45.82197547114778%**.
- Projections of original data points on principal components w1 and w2 can be obtained from $x(proj)_i = (x_i^T w_1)w_1 + (x_i^T w_2)w_2$ for all i.
- Projected and initial data points are displayed in the graph in the code file.

Justification for the assumptions in the above process:

- By pythagoras theorem, $||x_i||^2 = ||x_i - (x_i^T w)w||^2 + ||(x_i^T w)w||^2$

  $\frac{1}{n}\sum_{i=1}^{n}||x_i||^2$ = Average error + $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$

  Average error = $\frac{1}{n}\sum_{i=1}^{n}||x_i||^2$ - $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$

  First term ($\frac{1}{n}\sum_{i=1}^{n}||x_i||^2$) is not dependent on w which maintains its value consistent irrespective of the dataset. Therefore, in order to reduce error, we need to maximize the other term ($\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$).

  $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2 = \frac{1}{n}\sum_{i=1}^{n}(w^T x_i x_i^T \mathrm{w}) = w^T(\frac{1}{n}\sum_{i=1}^{n}x_i x_i^T)\mathrm{w} = w^T \mathrm{Cw}$,

  as $\mathrm{C} = \frac{1}{n}\sum_{i=1}^{n}x_i x_i^T = \frac{1}{n}(\mathrm{X}X^T)$.

  Now maximizing $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$ is same as maximizing w^TCw where w^Tw = 1.

  w_k are the principal components and are obtained from eigenvectors of the covariance matrix.

  $\mathrm{C}w_k = \lambda_k w_k$

  $w_k^T \mathrm{C} w_k = \lambda_k w_k^T w_k = \lambda_k$

  Covariance matrix is real symmetric. These eigenvalues obtained from covariance matrix are real and will be equal to the term $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$.

  Eigenvectors of the covariance matrix will be the principal components forming an orthogonal basis $\{w_1, w_2,..w_k\}$ with w_k being the eigenvector of the k-th largest eigenvalue i.e., k-th round.

- Variance of projected vectors = $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w - \mu)^2$, where mean $\mu = \frac{1}{n}\sum_{i=1}^{n}x_i^T w = 0$ as the data will be centered to the mean

Therefore, variance = $\frac{1}{n}\sum_{i=1}^{n}((x_i^T w)^2 + \mu^2 - 2\mu x_i^T w)$

$= \frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$ = λ = eigenvalue of the covariance matrix.

II. Without centering the data, values observed were almost the same with very minute changes in the order of e^{-8}.

- Calculated eigenvalues and eigenvectors of matrix C are:
    - λ_1 = 17.13191440244448 with corresponding eigenvector (-0.323516, -0.9462227) = w1(principal component 1)
    - λ_2= 14.489604749330741 with corresponding eigenvector (-0.9462227, 0.323516) = w2(principal component 2)
- Eigenvalues obtained are the variance of the corresponding principal component.
    - Variance corresponding to the principal component pc1 [-0.323516, -0.9462227] is **17.13191440244448.**
    - Variance corresponding to the principal component pc2 [-0.9462227, 0.323516] is **14.489604749330741**.
- Variance obtained from pc1 accounts for **54.178024528852205%** and that obtained from pc2 accounts for **45.82197547114779%** which is the same as the values obtained with centering (pc1 - **54.17802452885222%** ; pc2 - **45.82197547114778%**).

Centering doesn't make any difference here. Because in PCA we do eigen decomposition on a covariance matrix. It will find the axis in which direction the maximum data is spread. Covariance matrix remains the same with and without centering, therefore, we will always get the axis with maximum data spread.

III. Kernel PCA algorithm was implemented on the dataset with the given 2 kernels for all mentioned d and σ values. Projection of each point on the top 2 components for each kernel are plotted and compared in the code file.

Process followed:

- With the given kernel function, Kernel matrix was calculated and centered.

- Eigenvalues and eigenvectors for the centered kernel are calculated.
- Normalized this pair to obtain alpha
- Then mapped data points using centered kernel and alpha

IV. According to the observations made, Radial Basis Function (kernel B) seems to better suit the given dataset. While analyzing projected data points, we can see that the projections obtained by kernel B are linearly separable. Such formation of separate clusters helps us to easily analyze further tasks on them as performed in Q2 where we chose σ=3.8 as it gave 4 separations which are linearly separable and easy to form clusters using Llyod's algorithm.

It has also been observed that as sigma increased in Radial Basis Function, projections were more clearly separated which lets us to linearly cluster them.

2. Given a dataset with 1000 data points each in R^2

I. Functions assumemean, cluster, newmean, final, plotting are defined in code file to perform the algorithm with 4 clusters. In each initialization, a random data point is chosen and means are calculated with the functions defined above. Error function vs iterations and cluster plots are visualized in the code file.

II. Another function final_rand is defined here. We randomly allot initial cluster numbers to each datapoint here and then perform k-means to obtain final clusters. Clustering is done for given values of K and voronoi regions are plotted for each cluster center in the data file.

III. Spectral clustering was analyzed with all kernels and radial basis kernel with sigma 3.8 gave the best results compared to all. This is because the Radial Basis Kernel gave the dataset in 4 separations which are linearly separable yet Lloyd's algorithm failed to cluster this data. This is visualized in the code file with plots. Actual expected outcome is depicted by using kernel PCA and Lloyd's algorithm in the code file to show how H* matrix is unsuccessful in clustering the linearly separable data.

IV. By assigning data points directly to cluster l where l = arg $max_{j=1,2,...,k} v_i^j$ where v^j is the eigenvector of Kernel matrix associated with j-th largest eigenvalue, we are getting clusters directly from H* without further processes of normalizing and Lloyd’s algorithm. The clusters were visualized in the code file and the results are not as good as those obtained by spectral clustering.