Weekly report of lessons

Name: Suryam Arnav Kalra
Roll No: 19CS30050
The week: 4th October 2021 to 8th October 2021

The topics covered:

- Two major approaches for dimension reduction
- Principal component analysis (PCA)
- Computation of 1st component and 2nd component
- PCA algorithm, properties and application
- Linear discriminant analysis and fisher linear discriminant
- Separation between projected data of different classes
- A better measure of separation and scatter matrix

Summary topic wise:

- Two major approaches for dimension reduction: Feature selection => We find k of the d dimensions that give us the most information discarding the other dimensions. Feature extraction => We create a new set of k dimensions that are combination of the original d dimensions.
- Principal component analysis (PCA): It finds a mapping from the inputs in the original d-dimensional space to a new (k < d)-dimensional space, with minimum loss of information.
    - The criterion to be maximized is the variance. The principal component is w_1 such that the sample, after projection on to w_1, is most spread out so that the difference between the sample points becomes most apparent. Similarly, the other dimensions are given by the eigenvectors with **decreasing eigenvalues**.
    - The k eigenvectors with nonzero eigenvalues are the dimensions of the reduced space.
    - The first eigenvector (the one with the largest eigenvalue), w_1, namely, the principal component, explains the largest part of the variance; the second explains the second largest; and so on.
- Computation of 1st component: If $z_1 = {w_1}^T x$ with Cov(x) = Σ, then Var(z_1) = ${w_1}^T \Sigma w_1$ We seek w_1 such that Var(z_1) is **maximized** subject to the constraint that ${w_1}^T w_1$.The principal component is the eigenvector of the covariance matrix of the input sample with the **largest eigenvalue.**
- Computation of 2nd component: The second principal component, w_2, should also maximize variance, be of unit length, and be orthogonal to w_1. This latter requirement is so that after projection $z_2 = {w_2}^T x$ is uncorrelated with z_1.We get that w_2 should be the eigenvector of Σ with the **second largest eigenvalue**.
- PCA algorithm: We compute the mean of data points and translate all data points to their mean. Then, compute covariance matrix of the set and then compute the eigen vectors and the eigen values (in increasing order). Choose k such that the fraction of variance accounted for is more than a threshold.

- PCA properties: PCA diagonalizes the covariance matrix Σ. The components are uncorrelated since the covariance among the components is zero. By normalizing the components with their variances, Euclidean distance could be used for classification.
- Application of PCA: It is used in data compression, decorrelating components (eg: color images in RGB space are highly correlated), factor analysis and high level processing.
- Linear discriminant analysis: It captures the direction of maximum variance for a data set since for the purpose of classification, dimensional reduction using PCA may not work.
- Fisher linear discriminant: It is a projection of data x_i on a line with direction u as $y_i = {x_i}^T u$ which is a one dimensional subspace representing the data.
- Separation between projected data of different classes: Let m_1 be the mean of data points in w_1 and m_2 be the mean of data points in w_2 and let their projections be $m_{y1}\ and\ m_{y2}$ then $D = |m_{y1} - m_{y2}|$ is a measure of separation.
- A better measure of separation: We can normalize by a factor proportional to the class variances. $s^2 = \sum(y - m_c)^2$ and $J(u) = \frac{D^2}{({s_1}^2 + {s_2}^2)}$ where we want to obtain u which maximizes J(u).
- Scatter matrix: The scatter matrix for samples of class C in original space: $S_C = \sum(x - m_c)(x - m_c)^T$ and let $S_w = S_1 + S_2$ then ${s_1}^2 + {s_2}^2 = u^T S_w u$ and let $S_B = (m_1 - m_2)\ (m_1 - m_2)^T$ then $J(u) = \frac{u^T S_B u}{u^T S_w u}$

Concepts challenging to comprehend:

Linear discriminant analysis is a little bit challenging to comprehend.

Interesting and exciting concepts:

Principal component analysis is quite interesting and exciting to learn.

Concepts not understood:

After going through the book and the video lectures the concepts are understood.

Any novel idea of yours out of the lessons:

PCA is a strong technique which can be used in data processing. The data we collect can have many dimensions indicating different properties of the data. If we were to find the correlation between these different attributes, it will be highly expensive. In this case, we can use PCA to obtain the K best dimensions which represent the data and obtain the same result in a much less expensive way.

Difficulty level of the Quiz:

The quiz was somewhat on the tougher side.

Was the time given to you for solving the quiz appropriate? If not, why?

The time was almost appropriate.

Did the quiz questions enhance your understanding of the topics covered?

Yes, they helped me enhance my understanding of topics like DBSCAN.