PGP in AI/ML

Feature Engineering –Project-2

Submission Date: 23 59hrs on 11-08-2019

Total Marks: 12

In this Project the aim is to analyze breast cancer wisconsin dataset from sklearn & perform PCA(Principal Component Analysis) on this dataset.

1. Load the breast cancer dataset from sklearn datasets package. [1 M]
2. Store data & it's class in separate numpy arrays. [1 M]
3. Perform PCA on all 30 dimensions of data using PCA functions in sklearn decomposition package. [2 M]
4. Plot a graph with cumulative explained variance against number of components. [1.5 M]
5. By analyzing this graph find value of reduced number of components such that cumulative variance is maximized & number of principal components is minimized. [1 M]
6. Calculate explained_variance, explained_variance_ratio & cumulative sum of variance. [1.5 M]
7. Plot a scatterplot for top 2 principal components such that each class is represented by a separate color. Verify if the classes benign & malignant are separable in plot or not. [1 M]
8. Plot a 3D scatterplot for top 3 principal components, also plot two separate 2D plots among PCA1, PCA3 & PCA2, PCA3. [3 M]

sklearn PCA documentation for reference:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Submission Details

1. Python code file named YourName_PCA.py, code must print/output values mentioned in question 6.
2. Five PCA plots including three 2D plots, one 3D plot & one cumulative explained variance plot.
3. A document file named YourName_Project-2.doc containing all plots & calculated values of explained_variance, explained_variance_ratio & cumulative sum of variance.

Contact Details

You should put up queries in the discussion forum of the corresponding assignment folder only.

1. Himanshu - himanshuchawla437@gmail.com
2. Mohit - mohitrathore15@gmail.com
3. Kartheek - f20160015@hyderabad.bits-pilani.ac.in
4. Shristy - f20160688@hyderabad.bits-pilani.ac.in
a) Agam - f20160484@hyderabad.bits-pilani.ac.in