greatlearning

Unsupervised Learning Project

Data Description:

The data contains features extracted from the silhouette of vehicles in different angles. Four "Corgie" model vehicles were used for the experiment: a double decker bus, Cheverolet van, Saab 9000 and an Opel Manta 400 cars. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.

Domain:

Object recognition

Context:

The purpose is to classify a given silhouette as one of three types of vehicle, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles.

Attribute Information:

- All the features are geometric features extracted from the silhouette.
- All are numeric in nature.

Learning Outcomes:

- Exploratory Data Analysis
- Reduce number dimensions in the dataset with minimal information loss
- Train a model using Principle Components

Objective:

Apply dimensionality reduction technique – PCA and train a model using principle components instead of training the model using just the raw data.

Steps and tasks:

1. Data pre-processing – Perform all the necessary preprocessing on the data ready to be fed to an Unsupervised algorithm (10 marks)
2. Understanding the attributes - Find relationship between different attributes (Independent variables) and choose carefully which all attributes have to be a part of the analysis and why (10 points)
3. Split the data into train and test (Suggestion: specify “random state” if you are using train_test_split from Sklearn) (5 marks)
4. Train a Support vector machine using the train set and get the accuracy on the test set (10 marks)
5. Perform K-fold cross validation and get the cross validation score of the model (*optional*)
6. Use PCA from Scikit learn, extract Principal Components that capture about 95% of the variance in the data – (10 points)
7. Repeat steps 3,4 and 5 but this time, use Principal Components instead of the original data. And the accuracy score should be on the same rows of test data that were used earlier. (hint: set the same random state) (20 marks)
8. Compare the accuracy scores and cross validation scores of Support vector machines – one trained using raw data and the other using Principal Components, and mention your findings (5 points)

References:

- Book on PCA
- Application of PCA for image compression