Tools for Data Science

Task 3 PCA

Dev Moxaj Desai
Seasons of Code
Web and Coding Club, IIT B

Question 1

Conttinuing the quest into unsupervised learning algorithms, we soon encounter PCA (Principal Component Analysis). Simply put it is an algorithm used to reduce the dimensionality (no. of features) of the data.

This algorithm is generally used when overfitting occurs (feature selection), when the relationship between the features is to be studied (applications in finance and science in general) or when an image is to compressed (mainly in computer vision).

For this task, use the **iris** dataset (code: `datasets.load_iris()`) and **normalise** the data, you can use `StandardScaler` from `sklearn` or write a function of your own.

1. Reduce the normalised data to 2 dimensions and visualise it using a scatter plot (don't forget the legend)
2. Use k means clustering (with $k = 3$) on the 2D data obtained by PCA and the normalised data. Do both of them give similar results?

Basic Resoruces

1. Scatter plot with a legend
2. PCA
   (a) Basic intuition: 14.1 to 14.6
   (b) Algorithm in detail (Do not read before watching the basic part)
   (c) Implementation: a simple implementation
3. Clustering with `sklearn`