Labb 3 rapport

The report must include the following:   
• A short summary of the basic principle of PCA, with around 150-200 words   
• A short summary of the K-means algorithm, with around 150-200 words   
• A picture to visualize the objects of WINE data based on PCA   
• A table to give the centroid and number of data points for each cluster, based on the three   
experiments of clustering (using 2, 5, 8 principal components respectively)

Summary of the basic principle of PCA

By lowering the dimensionality of high-dimensional data, the widely used statistical method Principal Component Analysis (PCA) helps to discover patterns. The fundamental idea behind PCA is to change the original variables into a new collection of uncorrelated variables called principal components that account for most of the variance in the data. The maximum variance is chosen for the first principal component, and the maximum variance is then chosen for each component that follows it, if they are orthogonal to the ones that came before them. As a result, a series of principal components emerges that can be used to distill the data into a more manageable set of variables. PCA can be used with many different types of data, including images, acoustic signals, financial data, and biological data. It is helpful for visualizing and examining complicated datasets, finding patterns and correlations between variables, and decreasing the dimensionality of data while preserving the most data feasible. Data compression, feature selection, and visualization are just a few of the many data analytic uses for PCA.

Summary of the K-means algorithm

K-means is a known unsupervised clustering algorithm that divides a dataset into k different, non-overlapping clusters for machine learning and data mining. The algorithm randomly selects k initial cluster centers (centroids), assigns each point in the dataset to the closest centroid, and then iteratively updates the centroids by calculating the mean of all points assigned to each cluster. This process results in the creation of k clusters. Until the centroids reach a stable solution or a predetermined number of iterations has been reached, this process is repeated. The K-means algorithm aims to reduce the total squared distance between each point and its designated centroid. It is an algorithm that is easy to use and effective for handling large datasets with many dimensions.

The limitations of K-means' is it needs a specified number of clusters k beforehand, which can be difficult to determine beforehand. Moreover, the algorithm can be sensitive to initial centroid. Which means it can result in different final cluster assignments.

Result of PCA

En bild som visar text, diagram, Graf, linje

Automatiskt genererad beskrivning

|  |  |  |  |
| --- | --- | --- | --- |
| **2 PCA** |  |  |  |
| **Cluster** | **PC 1** | **PC 2** | **# data points** |
| 1 | -0,04 | 1,74 | 67 |
| 2 | -2,74 | -1,16 | 50 |
| 3 | 2,29 | -0,96 | 61 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **5 PCA** |  |  |  |  |  |  |
| **Cluster** | **PC 1** | **PC 2** | **PC 3** | **PC 4** | **PC 5** | **# data points** |
| 1 | 2,26 | -0,83 | -0,04 | -0,15 | -0,03 | 65 |
| 2 | -0,13 | 1,8 | 0,24 | 0,1 | -0,03 | 62 |
| 3 | -2,72 | -1,13 | -0,24 | 0,06 | 0,07 | 51 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **8 PCA** |  |  |  |  |  |  |  |  |  |
| **Cluster** | **PC 1** | **PC 2** | **PC 3** | **PC 4** | **PC 5** | **PC 6** | **PC 7** | **PC 8** | **# data points** |
| 1 | -0,04 | 1,77 | 0,19 | 0,08 | 0,07 | 0,13 | -0,002 | -0,02 | 65 |
| 2 | 2,28 | -0,93 | 0,002 | -0,14 | -0,13 | -0,22 | 0,05 | 0,025 | 62 |
| 3 | -2,72 | -1,13 | -0,24 | 0,06 | 0,07 | 0,1 | -0,06 | -0,01 | 51 |