Machine Learning Assessment

Task A - Clustering

Download BBC sports dataset from the Cloud. This dataset consists of 737 documents from the BBC Sport website corresponding to sports news articles in five topical areas from 2004-2005. There are 5 class labels: athletics, cricket,football, rugby, tennis.

1. There are 3 files in the dataset corresponding to the feature matrix, the class labels and the term dictionary. You need to read these files in Python notebook and store in variables X, trueLabels, and terms.

2. Next perform K-means clustering with 5 clusters using Euclidean distance as similarity measure. Evaluate the clustering performance using adjusted rand index and adjusted mutual information. Report the clustering performance averaged over 50 random initializations of K-means
3. Repeat K-means clustering with 5 clusters using a similarity measure other than Euclidean distance. Evaluate the clustering performance over 50 random initializations of K-means using adjusted rand index and adjusted mutual information. Report the clustering performance and compare it with the results obtained in step 2
4. For clustering cases (Euclidean distance and the other similarity measure), visualize the cluster centres using Tag cloud using Python package WordCloud.

Task B - (Dimensionality Reduction using PCA/SVD

For the provided BBC sports dataset, perform PCA and plot the captured variance with respect to increasing latent dimensionality. What is the minimum dimension that captures (a) at least 95% variance and (b) at least 98% variance?