COSC 425/528 — Introduction to Machine Learning — Fall 2018

Project 2: Dimensionality Reduction & Clustering

Due Oct 22, 2017

In this project you will apply dimensionality reduction and clustering to visualize information about some universities.

General Guidelines:

The same general guidelines given for Project 1 apply to this project (and to all others unless otherwise stated). Go through the steps of the workflow as you did in Project 1. (Note that, because this is not a prediction problem, you do not need to split your data into training, validation, and testing subsets.)

Specifics for Project 2:

In the folder for Project 2 on Canvas, you will find the data files, `UTK-peers.xls`, `UTK-peers.xlsx`, and `UTK-peers.csv`, each of which gives 65 attributes for UTK and 56 other similar universities.[1] Use whichever file is more convenient. I have eliminated several columns that were mostly empty, but you will have to decide what to do about remaining illegal values and non-numeric attributes.

Part 1 of this project will use principal components analysis (PCA) for data visualization.

1. Make a data matrix containing just the numeric attributes that you intend to use.
2. Use a library SVD package to factor your data matrix and extract the singular values.
3. Plot a scree graph of the singular values, and plot the percentage of variance covered by the first k singular values vs. k. (The variance is the square of the singular value.) What is a good choice of k?
4. Write a function to reduce your data matrix to the first k PCs, where k is the best value you have determined in step (3). (Hint: To do this, use the first k columns of your $\mathbf{V}$ matrix. Note that the SVD package returns $\mathbf{U}$, Σ, and $\mathbf{V}^T$, since it factors $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$.)
5. Make a scatter plot of the first two PCs. You can improve your plot by annotating the points with the universities' numbers or names.

(continues on next page)

1 This data is from the IPEDS database, https://nces.ed.gov/ipeds/datacenter/

In **Part 2** of this project you will implement *k*-means clustering and apply it to the original data.

6. Implement a *k*-means clustering program. Your program should take as arguments *k* (the number of clusters) and the data matrix to be clustered. Report the number of iterations required for convergence.
7. Report figures of merit for your clustering, including:
   a) the minimal *intercluster* distance (distance between points in different clusters),
   b) the maximal *intracluster* distance (distance of distinct points within a cluster),
   c) the *Dunn index*, which is the ratio of minimal *intercluster* distance to the maximal *intracluster* distance (a bigger ratio is better). You can decide how to compute the inter- and intracluster distances.
8. Use your program to cluster the data in (1) above. Experiment with different numbers of clusters and decide which best captures the structure of the data.
9. Take your cluster assignments and use them to annotate, color, or otherwise distinguish the clusters in your scatter plot from Part 1.
10. Which other universities are in the same cluster as UTK?
11. Repeat steps (8)–(10), but use the data matrix that represents the data in terms of the number of PCs you selected in step (3). Compare to your previous results.
12. Repeat step (11), but use only the first two PCs.

13. For **COSC 528** (extra credit for 425): Implement the EM algorithm for Gaussian clusters and repeat step (12) with it (i.e., cluster the data using the first two PCs).
14. Extra credit for both 425 and 528: There is another data file, `IPEDS-big-trimmed.csv`, which contains information about over 2000 universities. Analyze it in a similar way.