COSC 425/528 — Introduction to Machine Learning — Fall 2017

Project 2: Dimensionality Reduction & Clustering

Due Oct 23, 2017

In this project you will apply dimensionality reduction and clustering to visualize child mortality rates.

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.

Specifics for Project 2:

In the folder for Project 2 on Canvas, you will find the data files, `under5mortalityper1000.csv` and `under5mortalityper1000.xlsx`, which record under-5 childhood mortality per 1000 births from the World Health Organization. Use whichever file is more convenient. The first attribute is the country's name, and the remaining attributes are the deaths per 1000 for the years 1800–2017.

1. **Part 1** of this project will use principal components analysis for data visualization.
2. Make a data matrix containing just the yearly estimates (i.e., omitting the country name).
3. Use a library SVD package to factor your data matrix and extract the singular values.
4. Plot a scree graph of the singular values, and plot the percent of variance covered by the first k singular values vs. k. What is a good choice of k?
5. Write a function to reduce your data matrix to the first k PCs, where k is the best one you have decided in step (4). (Hint: To do this use the first k columns of your **V** matrix.)
6. Make a scatter plot of the first two PCs.

7. In **Part 2** of this project you will implement k-means clustering and apply it to the original data.
8. 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.
9. Report figures of merit for your clustering, including the minimal *intercluster* distance (distance between points in different clusters), the maximal *intracluster* distance (distance of distinct points within a cluster), and 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.
10. 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.
11. Take your cluster assignments and use them to annotate, color, or otherwise distinguish the clusters in your scatter plot from Part 1.
12. Repeat steps (10)–(11), but use the data matrix that represents the data in terms of its PCs. Compare to your previous results.
13. Repeat step (12), but use only the first two PCs.

14. For **COSC 528** (extra credit for 425): Implement the EM algorithm for Gaussian clusters and repeat steps (10) to (13) with it.