Homework 2

Directions: Complete all exercises.

1. Consider the MovieLense data that is available in the *recommenderlab* package

   >data(MovieLense)
   >?MovieLense

   The data was collected through the MovieLens web site during a seven-month, and contains about 100,000 ratings (1-5) from 943 users on 1664 movies. See the help file on the data to understand how to best manipulate the object.

   Design and evaluate your own recommendation system based on the following principles:
   For each user i and each movie j they did not see, find the k most similar users to i who have seen j and then use them to infer the user i's rating on the movie. Handle all exceptions in a reasonable way and report your strategy if you did so; e.g., if you cannot find k users for some movie j, then take all users that have seen it.

   Create the system so that outputs a user's top ten recommendations. Demo it on 3 users.

2. Data released from the US department of Commerce, Bureau of the Census is available in R (see, data(state)).

   (a) Focus on the data Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area. Cluster this data using hierarchical clustering. Keep the class labels (region, or state name) in mind, but do not use them in the modeling. Report your detailed findings. ** You may have done this step in an earlier assignment.

   (b) Focus on the data Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area. Cluster this data using SOM. Keep the class labels (region, or state name) in mind, but do not use them in the modeling. Report your detailed findings. ** You may have done this step in an earlier assignment.

   (c) Describe some of the advantages between the two above approaches in the context of this problem, and more generally.

3. Consider the Iris data (>data(iris)).

   (a) Create a plot using the first two principal components, and color the iris species by class.

   (b) Perform k-means clustering on the first two principal components of the iris data. Plot the clusters different colors, and the specify different symbols to depict the species labels.

   (c) Use rand index and adjusted rand index to assess how well the cluster assignments capture the species labels.

   (d) Use the gap statistic and silhouette plots to determine the number of clusters.

   (e) Reflect on the results, especially c-d. What does this tell us about the clustering?