University of Washington Bothell
Computing & Software Systems

Course title:	Introduction to Machine Learning
Course number:	CSS 490 / 590
Term:	Winter 2012
Instructor:	Jeff Howbert

Exercises 6

Date assigned:	Feb. 18, 2012
Date due:	Feb. 23, 2012, 10:00 PM

IDM refers to our textbook, "Introduction to Data Mining", by Tan, Steinbach, and Kumar.

Question 1) compares the quality of clustering obtained with k-means and agglomerative hierarchical clustering, using several different measures of distance or linkage for each. You will apply the accompanying script, `clust.m`, to a synthetic dataset, `synthGaussMix.mat` (see links in Schedule). The only thing you need to change in the script to answer this Question are the values of `distanceType`, `linkType`, and `doKmeans`. I recommend you save the modified script and run it from the command line, rather than executing by highlighting + F9.

`synthGaussMix.mat` contains 1428 datapoints and 6 continuous features. The datapoints were generated randomly from a small number *g* of multivariate Gaussian distributions. The value of *g* is in the range 2 to 8. The *g* multivariate Gaussians represent the "natural" cluster structure of `synthGaussMix`. The Gaussians have different means and covariance matrices (i.e. various ellipsoidal shapes and orientations), and the 1428 datapoints sampled from them overlap significantly. The main challenge of Question 1) is to make an informed guess as to the value of *g* after applying a spectrum of clustering methods and quality metrics.

---

1) Modify `clust.m` to run each of these 5 clustering variants:

- k-means with squared Euclidean distance
- k-means with cosine distance
- hierarchical clustering with single linkage
- hierarchical clustering with complete linkage
- hierarchical clustering with average linkage

Each run will automatically generate clusterings with k ranging from 2 to 8. For the hierarchical clusterings, this is done by cutting the tree obtained at different levels, so as to create 2 to 8 partitions (see function `cluster()` in the script).

For each value of k, two unsupervised measures of cluster quality are reported:

- SSE
- correlation

Review the sections of the textbook and lecture slides on these measures, and be sure you understand how they might give insight as to the number of “natural” clusters in a dataset.

For each value of k, the number of datapoints in each cluster is also reported.

a) [5 points] Cut-and-paste the output from the 5 variants into your answers document, and clearly label which run is which.
b) [15 points] One of the 5 variants clearly failed to produce a meaningful clustering.
   - Which variant is this?
   - Describe the different way in which each of the three pieces of information reported (SSE, correlation, datapoints per cluster) indicates the failure.
   - What properties of this variant might explain the failure? (Review the textbook and lecture slides on how this variant works.)
c) [15 points] Looking at the trends in SSE and correlation for the other 4 variants, what is your best guess for the number of “natural” clusters in the dataset? Try to find a consensus across the 4 variants. Explain your answer carefully, describing how you used the information in SSE and correlation to make your guess.

---

2) [5 points each, except 10 points for part (a)] Exercises 17(a), 17(b), 17(c), 17(d), and 17(f) in Section 8.7 of IDM (p. 563). Show your work and explain your answers.