CS 540-1: Introduction to Artificial Intelligence
Homework Assignment # 7

Assigned: 10/30
Due: 11/8 before class

Hand in your homework:

This homework includes only written questions. Please submit a single **hw7.pdf** file – we recommend latex, but you can use anything that can render nice math. Please typeset your homework, do not submit handwritten+scan answers. Include key steps. Go to UW Canvas, choose your CS540-1 course, choose Assignment, click on Homework 7: this is where you submit your files.

Question 1: PCA [25 points]

Consider a data set with 20 two-dimensional points of the form $(i, i), (i, i+1)$ for $i = 1, \ldots, 10$.

1. Center the data set by showing the new coordinates of (i, i) and $(i, i+1)$.
2. Compute the sample covariance matrix of the centered data $S = \frac{1}{n-1} X^\top X$, where X is the 20×2 matrix for the centered data.
3. Find the two principal components v_1, v_2. Each should be a two-dimension vector with norm 1. Hint: you may perform eigendecomposition online with `http://www.bluebit.gr/matrix-calculator/`. Here is an example to interpret the result. If the input matrix is (this is just an example, not your S)

   ```
   28.000 13.856
   13.856 12.000
   ```

   the program will return

   ```
   Eigenvalues:
   (36.000, 0.000i)
   ( 4.000, 0.000i)

   Eigenvectors:
   ( 0.866, 0.000i) (-0.500, 0.000i)
   ( 0.500, 0.000i) ( 0.866, 0.000i)
   ```

   Ignore the imaginary part. The format means the leading eigenvalue is 36, the corresponding eigenvector is the first column, namely $(0.866, 0.500)^\top$; the second eigenvalue is 4, and the corresponding eigenvector is $(-0.500, 0.866)^\top$.
4. Compute the new two-dimensional coordinates of CENTERED points – corresponding to the old (i, i) and $(i, i+1)$ – by projecting them onto v_1, v_2. (Hint: if you only keep the first dimension, it would be a maximum spread projection of the original data set.)

Question 2: Resolution [25 points]

Given the knowledge base

$$p \implies (q \implies r)$$

use resolution to prove the query

$$(p \wedge q) \implies (q \implies r).$$

Be sure to show what you convert to CNF and how (do not skip steps), and how you perform each resolution step.

Question 3: Hierarchical Clustering [25 points]

Consider the following six major cities. In the US: Madison, Seattle, Boston; and in Canada: Vancouver, Winnipeg, Montreal. For the purpose of this question ignore the curvature of the Earth, and compute the Euclidean distance. Suppose the cities are located at the following coordinates:

city	coordinate
Madison	(-89, 43)
Seattle	(-122, 48)
Boston	(-71, 42)
Vancouver	(-123, 49)
Winnipeg	(-97, 50)
Montreal	(-74, 46)

1. Use hierarchical clustering with complete linkage to produce TWO clusters by hand. Specifically, show the following in each iteration: (1) the closest pair of clusters; (2) the distance between them as defined by complete linkage; (3) all clusters at the end of that iteration.

2. Now repeat the above question, but with the following constraint: at no point should a US city and a Canadian city be put in the same cluster. Equivalently, whenever the complete linkage between two clusters is due to two cities in different countries, treat the two clusters as infinity apart, regardless of what other cities are in those two clusters. Show the same (1)(2)(3) as above in each iteration.

Question 4: K-means Clustering [25 points]

Given the following six items in 1D: $x_1 = 10, x_2 = 8, x_3 = 6, x_4 = 4, x_5 = 3, x_6 = 2$, perform k-means clustering to obtain $k = 2$ clusters by hand. Specifically,

1. Start from initial cluster centers $c_1 = 0, c_2 = 9$. Show your steps for all iterations: (1) the cluster assignments $y_1, \ldots, y_6$; (2) the updated cluster centers at the end of that iteration; (3) the energy at the end of that iteration.

2. Repeat the above but start from initial cluster centers $c_1 = 8, c_2 = 9$.

3. Which k-means solution is better? Why?