1. K-means Clustering

1) Given the following four vectors:
   \[ x_1 = <4, 1> \]
   \[ x_2 = <4, 3> \]
   \[ x_3 = <6, 2> \]
   \[ x_4 = <8, 8> \]

   Show how k-means clustering would cluster these values. Show how the cluster centers are moved on each iteration. Use the similarity function:
   \[ s(x_i, x_j) = -|x_i[1] - x_j[2]| - |x_i[2] - x_j[2]|, \]
   where \( x_i[1] \) represents the first component of the vector \( x_i \), etc. Assume that the cluster centers are initialized to \( <3, 2> \) and \( <7, 3> \).
1. **K-means Clustering**

\[\text{sim}(x_1, c_1) = -2, \quad \text{sim}(x_1, c_2) = -5\]
\[\text{sim}(x_2, c_1) = -2, \quad \text{sim}(x_2, c_2) = -3\]
\[\text{sim}(x_3, c_1) = -3, \quad \text{sim}(x_3, c_2) = -2\]
\[\text{sim}(x_4, c_1) = -11, \quad \text{sim}(x_4, c_2) = -6\]

\[c_1 = \left(\frac{4 + 4}{2}, \frac{1 + 3}{2}\right) = (4, 2)\]
\[c_2 = \left(\frac{6 + 8}{2}, \frac{2 + 8}{2}\right) = (7, 5)\]
1. *K*-means Clustering (Continued)

\[ c_1 = \left( \frac{4 + 4 + 6}{3}, \frac{1 + 3 + 2}{3} \right) = (4.67, 2) \]

\[ c_2 = \left( \frac{8}{1}, \frac{8}{1} \right) = (8, 8) \]

assignments remain the same, so the procedure has converged
2. EM Clustering

Consider a one-dimensional clustering problem in which the data given are:

\[
\begin{align*}
  x_1 &= -4 \\
  x_2 &= -3 \\
  x_3 &= -1 \\
  x_4 &= 3 \\
  x_5 &= 5 \\
\end{align*}
\]

Show the first two steps in the EM-based clustering procedure in which we model the clusters using two Gaussians. The initial mean of the first Gaussian is 0 and the initial mean of the second is 2. The Gaussians have fixed width; you should use EM to determine only the means. The density function to use for the Gaussians is:

\[
f(x, \mu) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{2} \right)^2}
\]

where \( \mu \) denotes the mean (center) of the Gaussian.

Note: this density function is slightly different than the one I originally specified. Here I have made the Gaussians wider, and I have included the normalizing term before the \( e \) part.
2. EM Clustering

\[ f(x, \mu) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{2})^2} \]

- \[ f(-4, \mu_1) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2}(\frac{-4-0}{2})^2} = 0.0269 \]
- \[ f(-4, \mu_2) = 0.0022 \]
- \[ f(-3, \mu_1) = 0.0646 \]
- \[ f(-3, \mu_2) = 0.00874 \]
- \[ f(-1, \mu_1) = 0.176 \]
- \[ f(-1, \mu_2) = 0.0646 \]
- \[ f(3, \mu_1) = 0.0646 \]
- \[ f(3, \mu_2) = 0.176 \]
- \[ f(5, \mu_1) = 0.00874 \]
- \[ f(5, \mu_2) = 0.0646 \]
2. EM Clustering: E Step

\[ h_1 = \frac{f(x_1, \mu_1)}{f(x_1, \mu_1) + f(x_1, \mu_2)} = \frac{.0269}{.0269 + .0022} \]
\[ h_{21} = \frac{f(x_2, \mu_1)}{f(x_2, \mu_1) + f(x_2, \mu_2)} = \frac{.0646}{.0646 + .00874} \]
\[ h_{31} = \frac{.176}{.176 + .0646} \]
\[ h_{41} = \frac{.0646}{.0646 + .176} \]
\[ h_{51} = \frac{.00874}{.00874 + .0646} \]
\[ h_2 = \frac{f(x_1, \mu_2)}{f(x_1, \mu_1) + f(x_1, \mu_2)} = \frac{.0022}{.0269 + .0022} \]
\[ h_{22} = \frac{.00874}{.0646 + .00874} \]
\[ h_{32} = \frac{.0646}{.176 + .0646} \]
\[ h_{42} = \frac{.176}{.0646 + .176} \]
\[ h_{52} = \frac{.0646}{.00874 + .0646} \]
2. EM Clustering: M-step

\[ \mu_1 = \frac{\sum_i h_{1i} \times x_i}{\sum_i h_{1i}} = \frac{-4 \times .924 + -3 \times .881 + -1 \times .732 + 3 \times .268 + 5 \times .119}{.924 + .881 + .732 + .268 + .119} = -1.94 \]

\[ \mu_2 = \frac{\sum_i h_{2i} \times x_i}{\sum_i h_{2i}} = \frac{-4 \times .076 + -3 \times .119 + -1 \times .268 + 3 \times .732 + 5 \times .881}{.076 + .119 + .268 + .732 + .881} = 3.39 \]
3. Threading

3) Consider a simple threading problem in which we have a template with three segments (i, j, k), each of which includes three amino acids. For a given sequence there are three possible starting positions for each segment. Suppose that you are given the following values for the scores of the individual segments and the scores for segment interactions.

\[
g_1(i,2) = 5 \quad g_1(j,8) = 9 \quad g_1(k,13) = 3 \\
g_1(i,3) = 2 \quad g_1(j,9) = 7 \quad g_1(k,14) = 4 \\
g_1(i,4) = 8 \quad g_1(j,10) = 6 \quad g_1(k,15) = 1 \\
\]

\[
g_2(i,j,2,8) = 1 \quad g_2(j,k,8,13) = 7 \quad g_2(i,k,2,13) = 1 \\
g_2(i,j,2,9) = 2 \quad g_2(j,k,8,14) = 8 \quad g_2(i,k,2,14) = 2 \\
g_2(i,j,2,10) = 2 \quad g_2(j,k,8,15) = 7 \quad g_2(i,k,2,15) = 5 \\
g_2(i,j,3,8) = 5 \quad g_2(j,k,9,13) = 1 \quad g_2(i,k,3,13) = 5 \\
g_2(i,j,3,9) = 6 \quad g_2(j,k,9,14) = 6 \quad g_2(i,k,3,14) = 6 \\
g_2(i,j,3,10) = 4 \quad g_2(j,k,9,15) = 8 \quad g_2(i,k,3,15) = 4 \\
g_2(i,j,4,8) = 7 \quad g_2(j,k,10,13) = 11 \quad g_2(i,k,4,13) = 1 \\
g_2(i,j,4,9) = 3 \quad g_2(j,k,10,14) = 12 \quad g_2(i,k,4,14) = 2 \\
g_2(i,j,4,10) = 4 \quad g_2(j,k,10,15) = 13 \quad g_2(i,k,4,15) = 4 \\
\]

Show how the branch and bound method would find the optimal threading. Show the threading sets considered in the search process and the lower bound calculated for each. Use the "simple lower bound" presented in class. When splitting a threading, split the segment having the minimal \( g_1 \) value (i.e. split on \( k \) first in the example above). To split the selected segment, divide it into three intervals of length one (this will work in the example because each segment has only three possible start positions).
3. Threading

\[ T = [2, 4], [8, 10], [13, 15] \]

\[ L_B = g_1(i, 3) + g_2(i, j, 2, 8) + g_2(i, k, 2, 13) + 
  g_1(j, 10) + g_2(j, k, 9, 13) + 
  g_1(k, 13) = 14 \]

\[ T = [2, 4], [8, 10], [13] \]

\[ L_B = g_1(i, 3) + g_2(i, j, 2, 8) + g_2(i, k, 2, 13) + 
  g_1(j, 10) + g_2(j, k, 9, 13) + 
  g_1(k, 13) = 17 \]

\[ T = [2, 8], [13] \]

\[ L_B = g_1(i, 2) + g_2(i, j, 2, 8) + g_2(i, k, 2, 13) + 
  g_1(j, 10) + g_2(j, k, 9, 13) + 
  g_1(k, 13) = 26 \]

\[ T = [3, 8], [10], [13] \]

\[ L_B = g_1(i, 3) + g_2(i, j, 3, 10) + g_2(i, k, 3, 13) + 
  g_1(j, 10) + g_2(j, k, 9, 13) + 
  g_1(k, 13) = 21 \]

\[ T = [3, 8], [10], [13] \]

\[ L_B = g_1(i, 3) + g_2(i, j, 3, 10) + g_2(i, k, 3, 13) + 
  g_1(j, 10) + g_2(j, k, 9, 13) + 
  g_1(k, 13) = 21 \]

\[ T = [4, 8], [10], [13] \]

\[ L_B = g_1(i, 4) + g_2(i, j, 4, 9) + g_2(i, k, 4, 13) + 
  g_1(j, 10) + g_2(j, k, 9, 13) + 
  g_1(k, 13) = 22 \]

\[ T = [2, 9], [13] \]

\[ L_B = g_1(i, 2) + g_2(i, j, 9, 2, 9) + g_2(i, k, 2, 13) + 
  g_1(j, 9) + g_2(j, k, 9, 13) + 
  g_1(k, 13) = 19 \]

\[ T = [2, 10], [13] \]

\[ L_B = g_1(i, 2) + g_2(i, j, 2, 10) + g_2(i, k, 2, 13) + 
  g_1(j, 10) + g_2(j, k, 10, 13) + 
  g_1(k, 13) = 28 \]
5. Parsimony

5) Given the following three sequences, draw the possible rooted trees for the sequences, and use Fitch's algorithm (traditional parsimony) to determine the hypothetical sequences for each internal node in the trees. Which tree would the parsimony principle prefer?

CGC  
TAT  
CAT
5. Parsimony

- three possible trees
- three sequence positions to consider
5. Parsimony (Continued)

3 total changes
5. Parsimony (Continued)

![Tree Diagram]

- CGC
- CAT
- TAT

{C}{GA}{CT}

2 changes

{CT}{A}{T}

1 change

3 total changes
5. Parsimony (Continued)

```
1  to  total  changes

TAT  CAT  CGC

1  

{CT}{A}{T}  {C}{AG}{TC}

2

3 total changes
```
5. Parsimony (Continued)

- so parsimony does not have a preference among these trees