

Total points: 75. Total time: 75 minutes. Answer all questions.
(This sample only has three of the questions from the mid-term exam of Fall 2007.
These are worth a total of 26 points, so about a third of the total worth of that exam.
Note that question 2 is from a section of the text that we have not covered, and is not
included in our syllabus this semester.)

Question 1. (a) (8 points) Consider a data set of five points in a 2D space:

A=(0,1)

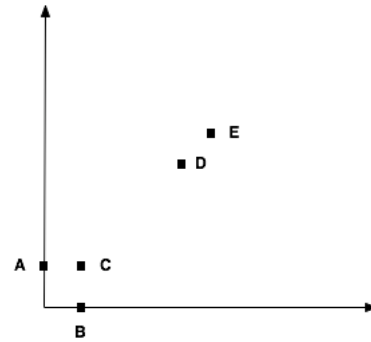
B=(1,0)

C=(1,1)

D=(4,4)

E=(5,5)

Consider the Lloyd algorithm for k-means clustering of this data set, with $k=2$. Suppose the current “means” or “centers” are: $M_1 = (0,0)$ and $M_2 = (1.5,1.5)$. Show what happens in the next two steps of the algorithm.



(b) (3 points) A common application of clustering methods is the clustering of “microarray data”. What does each data point in such an application correspond to? Each data point is typically a point in some high-dimensional space. What do these dimensions correspond to?

Question 2. (5 points) What is the theoretical mass spectrum of a peptide, for a given set of possible ion losses? What is the Shared-Peak-Count method for peptide sequencing?

Question 7. Consider the following instance of the exon chaining problem, where you are given 9 intervals, each with its own weight. The intervals (putative exons) are the line segments with square boxes at their end points, and the number sitting on each segment is the weight of that interval.

- (a) (4 points) Convert this problem into its graph representation (as discussed in class).
- (b) (6 points) Use the DAG you constructed to fill in the dynamic programming table associated with this problem. Report the score of the highest scoring (maximum weight) exon chain thus computed. **(You do not need to report the actual maximum weight exon chain.)**

