Due date: Thursday, October 13 @ 11:59

Accepted late without penalty: pretty much any time before the start of C term. At that point I become extremely busy again, and won’t be able to look at it.

You will work on expectation maximization in the context of analyzing flips from an unknown set of coins. **Note**: this problem is somewhat different than the one in the quiz. There are an unknown number of coins, each of which has a different probability of coming up heads. The algorithm to generate the data is:

For sequence\_num = 1 to max\_sequences

1. Select which coin to use at random
2. Flip that coin 10 times and record the outcomes (# heads and # tails).

Note that step (i) is not necessarily equiprobable. For example, P(coin 1) = 0.2, P(coin 2) = 0.3, P(coin 3) = 0.5. Also note that you’re not sure how many coins there are.

Your goal is to write a program that takes as input coin flip sequences of length 10 (see example file) and return:

1. The number of coins, *num*, used to generate the data set
2. P(heads | coini), for all i<num.

Notes:

1. You might want to generate a data set of your own with known numbers to help with debugging.
2. I would start by getting EM working with some fixed number of coins. Confirm that the basic algorithm is working. Using just one starting point is not normally acceptable, but is OK for this assignment.
3. Think about how you will determine how many coins there are. You can google on how to do EM with an unknown number of clusters. A simple approach is to put the algorithm you developed for note (2) inside of a loop, and test increasing numbers of clusters. A few points:
   1. The data will contain at most 10 different coins.
   2. Clustering with more cluster centers will provide a better model fit. Is there an inflection point where adding more cluster centers stops providing much more information?
4. The parameters for the sample coin flips file is:
   1. # coins = 3
   2. P(coin 1) = 0.2, P(coin 2) = 0.5, P(coin 3) = 0.3
   3. P(H | coin 1) = 0.15, P(H | coin 2) = 0.4, P(H | coin 3) = 0.8

What to hand in:

1. Your code and instructions on how to run it.
2. A 500 word writeup on how you approached the problem.
3. The writeup should also include your answer for the unidentified test set.