Programming Assignment 6 – Basic Probability, Computing and Statistics 2015

Fall 2015, Master of Logic, University of Amsterdam

Submission deadline: Monday, October 12th, 9 a.m.

Note: if the assignment is unclear to you or if you get stuck, do not hesitate to contact Philip.

1 Expectation Maximization

As we already pointed out during the lecture, EM is the go-to algorithm for unsupervised learning problems. We use it to learn the mixture weights for mixture models. The concrete procedure for learning those weights depends on what mixture components we assume or what kind of latent data. In this exercise you will implement your first instance of EM in the simplest possible scenario.

2 The Task

You are given a data set which contains the number of heads for 1000 sequences of 100 coin flips each. We will assume 9 mixture components, namely binomial distributions with θ-parameters 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 respectively. You should run EM from initially uniform mixture weights for 20 rounds. To compute the probability of the data points given the parameters, you can use the BinomialDistribution class that we implemented in week 4. After each iteration of EM you should print the log-likelihood for the data set. This is a good debugging technique. Since EM increases the likelihood at each iteration, you know you have a bug when the log-likelihood does not increase. Note: in later iterations, the likelihood may oscillate slightly and even decrease by small amounts. This has nothing to do with EM but with the numerical imprecision of your computer.

3 Grading

As is standard by now, your program should be runnable from the command line. The data file should then be provided as a command line argument.

2 points If the log-likelihood gets printed after each EM iteration.

2 points If there is a command line option that allows to set the number of EM iterations.

1 point If the specified number of iterations is not positive, the program should shut down gracefully (i.e. not crash), inform the user that non-positive iteration are not possible and print the help information from the command line parser.

2 points If the distribution over mixture components is printed after the final round of EM (notice the interaction with number of rounds here: the distribution should always be printed after the last round, no matter what the last round is. If the distribution is only printed after a specific number of rounds, give no points here).

3 points If EM is implemented correctly. This can be checked in two ways: first, the likelihood should keep increasing (modulo numerical imprecision in later rounds) and second, the final distribution should be correct.

For the purpose of grading, the log-likelihood and the distribution that you should compute after 20 iterations will be made available next week. Notice that those have been computed on our machines and yours may yield slightly different results. However, the numbers should be in the same ball park.