CS425
Assignment 3 (Due Mar 9, 2017)
Hidden Markov Models

Computer Science Department
Colorado State University

February 27, 2017

1. Implementing the forward algorithm [30 pts].
As you have learned, each cytosine nucleotide of DNA can be either methylated or unmethylated; this happens in every cell. When we look at a population of millions of cells, the same location (nucleotide) can be methylated in a fraction of the cells and unmethylated in the rest. Hence, we assign a DNA methylation level between 0 and 1 to each genomic cytosine, based on the fraction of the cells in which that position has been methylated. For simplicity, we call the methylation level of a cytosine "low" if it is below 0.5 and "high" otherwise.

The genome is then partitioned into hypo- and hypermethylated regions. Cytosines have usually low methylation levels in the hypomethylated regions and high methylation levels in hypermethylated regions. We can think about hypo- and hypermethylated regions as two different states in an HMM, and the DNA methylation of a cytosine as the emission. Consider that in a hypomethylated region the probability that a cytosine has low/high methylation levels is 0.8/0.2 respectively. In a hypermethylated region those probabilities are 0.2 and 0.8.

Your task is to implement the HMM forward algorithm to compute the observation probability for the following sequence of consequent cytosine methylation levels:

$$0.1, 0.3, 0.2, 0.4, 0.7, 0.9, 1, 1, 0.8, 0.9$$

Consider the initial probability of each state is 0.5, and the transition probability from one state to another after an emission is 0.1; hence the probability of staying in the same state is 0.9. Since the observation probability would be too small, your program should print the ln of this probability. Your are welcome to see the other implementations, by your classmates or over the internet, but the final implementation you deliver should be completely yours.

2. Implementing the backward algorithm [30 pts].
Add the backward algorithm to your program to calculate the ln of the probability of the 5^{th} cytosine above to be inside a hypermethylated region.

3. Learning the HMM parameters from real data [40 pts].
Download the DNA methylation levels of mouse embryonic stem cells from the Gene Expression Omnibus database `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30202` (Look at

bottom of the page, Supplementary file, and download GSE30202_BisSeq_ES_CpGmeth.tsv.gz). It contains the chromosome number, position of each cytosine, the total number of next generation sequencing reads covering the cytosine, and the number of reads in which cytosine has been methylated.

Dividing the fourth column by the third one gives you the methylation level of each cytosine. Keep only the cytosines of sexual chromosome Y (chrY) in the initial order, assign them the low and high methylation levels, and learn a HMM using this sequence of observations to obtain more realistic emission and transition probabilities. For this purpose you may use any implementation of the Baum-Welch algorithm in any programming language package. For the initial HMM settings, use the model described above in task 1.

Upload your answer on Canvas in one zip file or tarball. Include all the code/scripts you have written in your submission.