

CSC 575: Interactive Machine Learning
Homework 1 (to be worked on individually)

Assigned: 25th of February, 2015

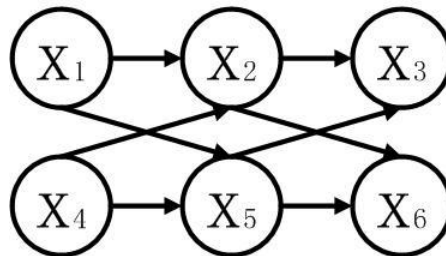
Due: 11th of March, 2015 (at the beginning of the class)

Total number of points: 100

1. For the following cases, is it appropriate to model their data using HMMs? Explain the reasons. (20 pts)

- Monthly precipitation data in Rochester
- A dataset of indoor images
- Handwriting recognition
- Daily S&P 500 stock market price

2. Follow the graph to decompose the joint probability $P(x_1, x_2, x_3, x_4, x_5, x_6)$. (40 pts)



3. A DNA sequence is a series of components $\{A, C, G, T\}$. Assume the hidden variable X takes 2 possible state values $\{s_1, s_2\}$, and the parameters of the HMM M are as follows: (40 pts)

Transition probabilities: $P(s_1|s_1) = 0.8, P(s_1|s_2) = 0.2$

$$P(s_2|s_1) = 0.2, P(s_2|s_2) = 0.8$$

Emission probabilities: $P(A|s_1) = 0.4, P(C|s_1) = 0.1, P(G|s_1) = 0.4, P(T|s_1) = 0.1,$

$$P(A|s_2) = 0.1, P(C|s_2) = 0.4, P(G|s_2) = 0.1, P(T|s_2) = 0.4,$$

Initial probabilities: $P(s_1) = 0.5, P(s_2) = 0.5$

The observed sequence is $Y = CGTCAG$.

- Calculate $P(Y|M)$. (hint: refer to the lecture slides to calculate forward messages recursively, each forward message is a vector containing the probabilities of the two hidden states at that time step)
- Calculate $P(x_3 = s_1|Y)$. (hint: refer to the lecture slides for the recursive derivation of the forward-backward algorithm)