CFRM 521 Machine Learning for Finance, Spring 2018.

Homework 3 and 4

Due: Sunday 29 April 2018, NO LATER than 11:59pm

Instructions: You must submit your work through Canvas, and in the form of two attachments: A .pdf file of your detailed answers and the R script of the codes you used. The .pdf files can be typeset or be a readable scan of your handwritten notes. You do not need to include R outputs in the .pdf file. However, you must clearly refer to the relevant lines of your R codes when answering a question that is based on your code.

The maximum number of points you can receive for this homework is 30.

1. Consider the setting of exercise 10.3 of the textbook [B] (Whizzco decide to make a text classifier...), which you have solved in homework 2. Now, assume that the training data has some missing values, as in the data file "HW3-Q1.txt" that accompanies this homework. Use the following code to load the data into R:

```
dat = read.table("HW3-4-Q3.txt",header=T)
```

   The missing data points are indicated by "NA" values.

   (a) (2 pts) Assume that the MAR assumption holds. Draw an appropriate belief network for the naive Bayes setting with missing data. Derive the marginal log-likelihood.

   (b) (4 pts) Derive the "E" and "M" steps of the EM algorithm.

   (c) (4 pts) Implement the EM algorithm in R, and use it to find the probability that the document $x = (1, 0, 0, 1, 1, 1, 1, 0)$ is about politics. Compare your results with what you found in Homework 2.

2. (a) (4 pts) In unprofitable times corporations sometimes suspend dividend payments. Suppose that, after a dividend has been paid, the next quarter dividend will be paid with probability 0.8, while after a dividend is suspended, the one for the next quarter will be suspended with probability 0.5. Also, we know that in the second quarter of 2015, 10% of the companies in the S&P500 index suspended their dividend payment. Assume that whether a company pays or retains the dividend can be modeled as a Markov chain. What percentage of the companies in the S&P500 index will pay dividend in the first quarter of 2016?

   (b) (6 pts) Solve exercise 23.3 of the textbook [B] (Consider an HMM with three states (M=3) and two output symbols...).

3. (a) (4 pts) Consider Example 23.2 of the textbook [B] (gene clustering), which is an application of mixture of Markov chains. Write an R code that solves this example. Check that your answers are consistent with the results in the textbook.

   (b) (4 pts) Let P be the transition matrix of the MC that generates the following sequence:

   A, C, G, T, A, C, G, T, A, C, G, T, A, C, G, T,...

   The "Randomize" modification of this MC is $P_{\text{new}} = 0.9\,P + 0.1[0.25]_{4\times4}$, where $[0.25]_{4\times4}$ is a 4×4 matrix with elements equal 0.25. Similarly, let $Q_{\text{new}} = 0.9\,Q + 0.1[0.25]_{4\times4}$, where Q is the transition matrix of the MC that generates the sequence

   T, G, C, A, T, G, C, A, T, G, C, A, T, G, C, A,....

What is the probability that the MC with transition P_{new} generated the following sequence?

A, A, G, T, A, C, T, T, A, C, C, T, A, C, G, C

What is the probability that the sequence was generated by the MC with transition Q_{new}?

(c) (2 pts) Generate 100 sequence of length 16 from the MC with transition P_{new}, and 100 sequence of length 16 from the MC with transition Q_{new}. Put the 200 sequence together to create a sample of size 200 of sequences of length 16. Use the R code that you wrote for part (a) to cluster this sample into two groups. Do you agree with the solution obtained by your code?
Hint: You may find the function "markovchainSequence()" in the R package "markovchain" useful.