CS769 Advanced NLP Homework 6

Due 4/21/2010

1 Hidden Markov Models

Andy is a three-month old baby. He can be happy, hungry, or having a wet diaper. Initially when he wakes up from his nap at 1pm, he is happy. If he is happy, there is a 50% chance that he will remain happy one hour later, a 25% chance to be hungry by then, and a 25% chance to have a wet diaper. Similarly, if he is hungry, one hour later he will be happy with 25% chance, hungry with 25% chance, and wet diaper with 50% chance. If he has a wet diaper, one hour later he will be happy with 50% chance, hungry with 25% chance, and wet diaper with 25% chance. When he is happy, he smiles 75% of the time and cries 25% of the time; when he is hungry, he smiles 25% and cries 75%; when he has a wet diaper, he smiles 50% and cries 50%.

Q1. **[6]** Draw the HMM (not the graphical model) that corresponds to the above story. Clearly mark the transition probabilities and output probabilities.

Q2. **[10]** The nanny left a note: "1pm: smile. 2pm: cry. 3pm: smile". What is the probability that this particular observed sequence happens?

Q3. **[6]** Draw the directed graphical model for the note in Q2. Then convert it into a factor graph. Clearly define each factor.

Q4. **[6]** What is the most likely hidden sequence (in terms of happy, hungry, or wet diaper) for the note in Q2? Use the max-sum (Viterbi) algorithm.

Q5. **[6]** Use the sum-product algorithm to compute the marginal state probabilities at 1pm, 2pm and 3pm, respectively.

Q6. **[6]** Based on the note in Q2, is Andy going to smile or cry at 4pm? Show your steps.

2 Information and Compression

I have a 4-sided die with probabilities $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ for faces 1,2,3,4, respectively. You know the probability distribution. I repeatedly roll the die and generate

outcomes $x_1, x_2, \ldots$. Your goal is to design an efficient "prefix code" binary encoding, represented as a binary tree, such that my outcomes are represented as a bit stream as short as possible.

Q7. [**6**]: Figure 2 shows a prefix code. The four outcomes are at the leaves, and the path from root is the encoding. For example, 2 is 01, 3 is 10. If we use this code, what is the *average number of bits per outcome* for my sequence $x_1, x_2, \ldots$?

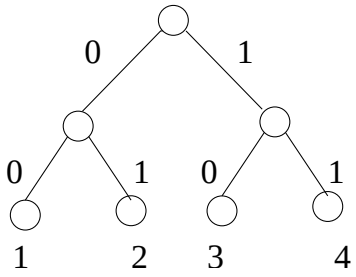

Figure 1: The prefix code for Question 1.

Q8. [**6**]: Draw the best binary tree encoding that you can think of. What is the *average number of bits per outcome* using your tree? Explain the connection to *entropy*.

Q9. [**6**]: Unknown to you and me, someone substituted my die with a fair die (equal probability) before my rolls. What is the *average number of bits per outcome* using your tree?

3 Feature Selection using Mutual Information

Sometimes people do not use the whole vocabulary as the set of feature. They select a subset of the features that are "more useful". In classification, this is often done by computing the mutual information between a word type and the class label.

In `http://pages.cs.wisc.edu/~jerryzhu/cs769/dataset/movie/`, there are two files. Each line is a movie review document. All documents in `positive.txt` have label $y = 1$, and in `negative.txt` have label $y = -1$. Do not further tokenize the files – treat each string separated by whitespace as a word. Punctuations are words too.

Q10. [**6**]: Create a *single* vocabulary from both positive and negative documents. You do not need to include special symbols like $\langle s \rangle$. What is the vocabulary size (i.e., the number of word types)?

Q11. [**6**]: We will treat each document as a random trial. The class label y is a random variable, and takes one of the outcomes -1, 1. *Each* word type w is a random variable too, which is 1 if w appears in the document, and 0 otherwise. Note we ignore the actual integer count of w in the document. Write down the formula for estimating the mutual information between y and w. Briefly explain what counts you need to collect for each y, w pair from the movie dataset.

Q12. [**6**]: Compute the mutual information for all pairs of y, w. Sort them. List the top 20 w's with the largest mutual information.

Q13. [**6**]: Does mutual information tell you if w is a "positive" or a "negative" word? If so, explain why. If not, suggest (informally) a way to obtain that information.

4 Spectral Clustering

Consider a graph with 10 items $x_1 \dots x_{10}$. All pairs within the first 5 items $x_1 \dots x_5$ are fully connected with edges of weight 1. All pairs within the last 5 items $x_6 \dots x_{10}$ are fully connected with edges of weight 1 too. Furthermore, x_1 and x_6 are connected with an edge of weight t.

Q14. [**6**] Show the weight matrix and the *unnormalized* Laplacian matrix.

Q15. [**6**] Show the smallest two eigenvalues of the Laplacian, and the corresponding eigenvectors, for $t = 1$. Interpret the results from a clustering point of view.

Q16. [**6**] Repeat Q15 for $t = 100$.