L&S II: Assignment 1

Prof. Noah Smith

Due: Thursday, September 14 (hardcopy, in class)

Since this course deals mainly with text, and the most commonly used text databases come from newspapers, this assignment deals with a newspaper route. Pat is a grad student who delivers newspapers for extra cash. In this exercise, you will try to model Pat's behavior as best you can using statistical models. The goal is to get you to think about how you can capture the observed phenomena using a model, and to get some exercise talking about models clearly.

Every Saturday, Pat delivers newspapers. These homes are distributed among several different neighborhoods in the suburban town of Slumberville. Because Pat gets bored easily, Pat likes to vary the route.

Let each house receive a numerical different code, from 0 to 999. These codes have nothing to do with Pat, who doesn't know about them; they are assigned by the Slumberville government for the purpose of maintaining zoning and land use records, sales of property, and so forth. They correspond loosely with the geographical layout.

Your job is to predict Pat's route in real time on a given week. For each problem, you will get a little more information about Pat's route that should help you build a better model. The goodness of your model will be measured by the probability it assigns to the route Pat chooses next week, which you won't see until the day before the deadline (i.e., this is the test data).

For each problem, you are to implement a computer program that, given Pat's route on a given week, outputs an estimate of the natural logarithm of the probability that Pat would take the observed route. You can (and should) use the data given to train a model for each problem. You are likely to need hidden variables to make use of the facts given in the problem set, and it's up to you how to deal with that during training and during testing. (Note that if you do not marginalize over hidden variables at test time, your log-likelihood scores will suffer!) Finally, you should be prepared to turn in your code if we have trouble believing in the quality of your estimates.

Pat's routes for the last year are available in `http://www.cs.cmu.edu/~nasmith/LS2.F06/a1.train.dat`, with one route per line.

1. Before doing anything else, answer this question: why are Markov models inherently unsuited for describing the distribution over Pat's bicycle routes? What is the price your model will have to pay if you keep the Markov property with a small Markov order?

2. Define and describe a probability model based on what you already know *and the data*. Implement a program whose input is a sequence of non-repeating whitespace-separated integers in $[0, 1000)$ and whose output is the natural logarithm of your model's probability of Pat following the given sequence. (It's useful to think for a second about a baseline answer to this exercise that doesn't use the data. If Pat delivers a newspaper to each of the 1,000 houses, then there are 1000! possible routes. The uniform model over those routes will assign log probability $-\ln(1000!) \approx -5912$ to each route. You should be able to do better!)

3. Pat doesn't live anywhere near Slumberville and usually chooses to take a bus to get there. The two bus stops that have direct lines from Pat's house into the suburb are located in front of houses 432 and 627. Pat can go home on bus lines that stop at the same bus stops. (He can take his bicycle on the bus.) Give the definition of your revised model and implement the scoring function as in (2).

4. Along the same side of a street, the house codes run consecutively (unlike street addresses, which in the US usually alternate). When a corner is turned or when walking across the street, the numbers may or may not be consecutive. Give the definition of your revised model and implement the scoring function as in (2).

5. Part of the suburb has grid-like streets that run north-south or east-west (where many streets are one-way), and part of the suburb has typical suburban streets that wind around in circles and frequently end in cul-de-sacs (where all streets are two-way). Pat prefers not to toss newspapers across a lane of oncoming traffic. Give the definition of your revised model and implement the scoring function as in (2).

6. Pat usually starts the route at 9am, takes a lunch break at a café near house 101 from 12–1pm (during which Pat reads research articles on computational linguistics), then continues the route until 4pm. Give the definition of your revised model and implement the scoring function as in (2).

Remember, in addition to answering each question concisely and clearly in English, you need to report values of the log-probability of the test data (available the day before the deadline on the course web page) for questions 2–6! Part of your grade will be based on how well this value compares with your classmates.

7. NLP Question Suppose you must build a language model for a language like Chinese that is not normally whitespace separated. You are given a corpus of unsegmented text; you do not have a word list. Part 1: brainstorm some solutions to this problem and discuss. Part 2: how would your answer change if you had access to a (partial) word list? If you know about this problem, or if you look for relevant literature when answering the question, you must cite prior work!

8. Optional Bonus Question Back to Pat—how would you go about inferring a geographical map of the suburb, given observations of Pat's route over a period of time?