

## CS 445/545: Machine Learning Winter, 2009

### Homework 3, Part 2: Due Wednesday March 11, 5:00pm.

#### 3. Hidden Markov Models “Madlibs”

In this exercise you will train a hidden Markov model from text data. The observations are the words in the text, and the hidden states are the parts-of-speech tags for each word. Here is a short example:

```
‘Ryan walked in the door.’
```

```
Observations: {'ryan' 'walked' 'in' 'the' 'door' '.'}
```

```
Hidden states: {'name' 'verb' 'preposition' 'article' 'noun' 'punctuation'}
```

Your task: Find a short “story” (a paragraph or two of text). Create a training file with the observations and hidden states (you’ll have to do the part-of-speech labeling for each word).

Run the Matlab function `hmmestimate` using your training data to create the hidden-state transition probability matrix and the observations “emission” probability matrix.

Then run the Matlab function `hmmgenerate` using these matrices to probabilistically generate a new “story” of about the same length as your original story. Repeat (you’ll get a different story).

Note that figuring out how to use these two Matlab functions from their documentation (in the Statistics toolbox) is part of this exercise. Feel free to discuss this with your classmates (and on the class mailing list) if you’re having problems.

**Optional:** Experiment with different sizes of training data to see how the “reasonableness” of the generated stories change with size of training data.

#### What to turn in

The original story, and two hmm-generated “stories”. Also the hidden-state transition matrix and the observation emission matrix.