ENGN 2520 / CSCI 1950-F Homework 5
Due Friday March 15 by 4pm

Problem 1

Let $S = \{s_1, \ldots, s_K\}$ be the set of states of a Markov chain. Let π be the initial state distribution and M be the transition matrix. Suppose we have a training set $T = \{x_1, \ldots, x_N\}$ consisting of independent samples (sequences) from the Markov chain. Sequence i has length n_i and is given by $x_i = (x_{i,1}, \ldots, x_{i,n_i})$.

    (a) Derive the MLE estimate for the initial state state distribution π and the transition matrix M? Justify your answer.

    (b) The rows of M are distributions over states. Row i defines $p(x_{t+1} = s_j | x_t = s_i)$. Suppose we use a dirichlet prior $\text{Dir}(\alpha)$ with a single parameter α for the initial state distribution π and each row of the transition matrix M. The dirichlet prior defines a prior probability for a multinomial distribution with parameters $\mu = (\mu_1, \ldots, \mu_K)$ as

$$\text{Dir}(\mu_1, \ldots, \mu_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \mu_i^{\alpha-1}$$

where $B(\alpha)$ is a constant that depends only on α.

    Derive the MAP estimate for π and M. How does it differ from the MLE?

Problem 2

Suppose we have an HMM with states $S = \{s_1, \ldots, s_K\}$, observations $O = \{o_1, \ldots, o_M\}$, initial state distribution π, transition matrix M and observation matrix H.

    Let $y = \{y_1, \ldots, y_n\}$ be a sequence of observations. Show how we can efficient compute the conditional probability $P((x_t = s_i), (x_{t+1} = s_j), (x_{t+2} = s_k) | y)$ using forward and backward weights.

Problem 3

In this assignment you will implement the Viterbi algorithm for correcting corrupted english sentences. For this assignment each english sentence is a sequence of characters (states), from

an alphabet (state space). The state space has 26 states corresponding to the 26 lowercase letters and the space character.

We will assume that each english sentences are generated by a first order Markov model. We observe each sentence through a noisy channel that corrupts each character with probability ϵ. If a character is corrupted it gets replaced by an arbitrary character with equal probability (a character can be replaced by itself). This process leads to a HMM. The class website contains examples of uncorrupted english sentences that you will use to train the parameters of the model.

(a) What is the state space for the hidden variables? What is the observation space?

(b) What is the observation matrix H in terms of ϵ?

(c) Use the uncorrupted sentences available on the class website to estimate the initial state distribution π and the transition matrix M. Use your model to generate 10 independent samples of length 50 from the joint distribution $p(x_1, \ldots, x_n, y_1, \ldots, y_n)$ when $\epsilon = 0.0, 0.1$ and 0.2.

(e) Implement the Viterbi algorithm for computing the most likely sequence of hidden states $(x_1, \ldots, x_n)$ given a sequence of observations $(y_1, \ldots, y_n)$. Use the algorithm to "correct" the corrupted sentences available on the class website using the model from part (d) when $\epsilon = 0.1$. You should turn in the results of correcting the first 10 sentences.