CSE 250A. Assignment 7

Out: *Tue Nov 8*
Due: *Tue Nov 15*

Supplementary reading:

- Russell & Norvig, Chapter 15.
- L. R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

7.1 Viterbi algorithm

In this problem, you will decode an English sentence from a long sequence of non-text observations. To do so, you will implement the same basic algorithm used in most engines for automatic speech recognition. In a speech recognizer, these observations would be derived from real-valued measurements of acoustic waveforms. Here, for simplicity, the observations only take on binary values, but the high-level concepts are the same.

Consider a discrete HMM with $n = 26$ hidden states $S_t \in \{1, 2, \ldots, 26\}$ and binary observations $O_t \in \{0, 1\}$. Download the ASCII data files from the course web site for this assignment. These files contain parameter values for the initial state distribution $\pi_i = P(S_1 = i)$, the transition matrix $a_{ij} = P(S_{t+1} = j | S_t = i)$, and the emission matrix $b_{ik} = P(O_t = k | S_t = i)$, as well as a long bit sequence of $T = 180000$ observations.

Use the Viterbi algorithm to compute the most probable sequence of hidden states conditioned on this particular sequence of observations. Turn in the following:

(a) a print-out of your source code

(b) a plot of the most likely sequence of hidden states versus time.

You may program in the language of your choice, but it will behoove you to consider the efficiency of your implementation in addition to its correctness. Well-written code should execute in seconds (or less).

To check your answer: suppose that the hidden states $\{1, 2, \ldots, 26\}$ represent the letters $\{a, b, \ldots, z\}$ of the English alphabet. The most probable sequence of hidden states (ignoring repeated letters) will reveal a timely message for this election season.