Notes for https://web.stanford.edu/~jurafsky/slp3/9.pdf

Algis Dumbris

Version 1.0

#:toc: left

Introduction

Nabokov famously translated it strictly literally Pushkin's novel Eugene Onegin into English prose.

In 1913, A. A. Markov asked a less controversial question about Pushkin's text: could we use frequency counts from the text to help compute the probability that the next letter in sequence would be a vowel?

Hidden Markov Model or HMM.

The HMM is a probabilictic sequence model. A sequence model or sequence classifier is a model whose job is to assign a label or class to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels.

Markov chain, sometimes called the observed Markov model, it is weighted finite automaton.

Markov chain is only useful for assigning probabilities to unambiguous sequences

3 problems

The hidden Markov models should be characterized by three fundamental problems:

- **Problem 1 (Likelihood):** Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O|\texttt{Unknown characterUnknown character})$.
- **Problem 2 (Decoding):** Given an observation sequence O and an HMM $\texttt{Unknown characterUnknown character} = (A, B)$, discover the best hidden state sequence Q.
- **Problem 2 (Learning):** Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

Likelihood computation, forward algorithm

But of course, we don't actually know what the hidden state (weather) sequence was. We'll need to compute the probability of ice-cream events 3 1 3 instead by summing over all possible weather sequences, weighted by their probability. First, let's compute the joint probability of being in a particular weather sequence Q and generating a particular sequence O of ice-cream events. In general, this is

$$P(O, Q) = P(O|Q) \times P(Q) = \prod_{i=1}^{T} P(o_i|q_i) \times \prod_{i=1}^{T} P(q_i|q_{i-1})$$

The computation of the joint probability of our ice-cream observation 3 1 3 and one possible hidden state sequence hot hot cold is

$P(313, hothotcold) = P(hot|start)\texttt{Unknown characterUnknown character}P(hot|hot)\texttt{Unknown characterUnknown character}P(cold|hot)\texttt{Unknown characterUnknown character}P(3|hot)\texttt{Unknown characterUnknown character}P(1|hot)\texttt{Unknown characterUnknown character}P(3|cold)$

We can compute the total probability of the observations just by summing over all possible hidden state sequences:

$P(O) = \sum_Q P(O, Q) = \sum_Q P(O|Q)P(Q)$

For our particular case, we would sum over the eight 3-event sequences cold cold cold, cold cold hot, that is, P(3 1 3) = P(3 1 3, cold cold cold) +P(3 1 3, cold cold hot) +P(3 1 3,hot hot cold) +…

Instead of using such an extremely exponential algorithm, we use an efficient $O(N^2T)$ algorithm called the forward algorithm. The forward algorithm is a kind Forward algorithm of **dynamic programming algorithm**, that is, an algorithm that uses a table to store intermediate values as it builds up the probability of the observation sequence.