CSE 250A. Assignment 6

Out: *Tue Nov 10*
Due: *Tue Nov 17*

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.

6.1 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 $S_t \in \{1, 2, \ldots, z\}$ 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 $T = 75000$ 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 hard-copy 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 the upcoming holiday season.

6.2 Inference in HMMs

Consider a discrete HMM with hidden states S_t, observations O_t, transition matrix $a_{ij} = P(S_{t+1}=j|S_t=i)$ and emission matrix $b_{ik} = P(O_t=k|S_t=i)$. In class, we defined the forward-backward probabilities:

$$\begin{aligned} \alpha_{it} &= P(o_1, o_2, \ldots, o_t, S_t=i), \\ \beta_{it} &= P(o_{t+1}, o_{t+2}, \ldots, o_T|S_t=i), \end{aligned}$$

for a particular observation sequence $\{o_1, o_2, \ldots, o_T\}$ of length T. This problem will increase your familiarity with these definitions. In terms of these probabilities, which you may assume to be given, as well as the transition and emission matrices of the HMM, show how to (efficiently) compute the following probabilities:

(a) $P(S_t=i|S_{t+1}=j, o_1, o_2, \ldots, o_T)$

(b) $P(S_{t+1}=j|S_t=i, o_1, o_2, \ldots, o_T)$

(c) $P(S_{t+1}=k|S_{t-1}=i, S_t=j, o_1, o_2, \ldots, o_T)$

(d) $P(S_{t+1}=k|S_{t-1}=i, o_1, o_2, \ldots, o_T)$

In all these problems, you may assume that $t > 1$ and $t < T$; in particular, you are *not* asked to consider the boundary cases.

6.3 Belief updating

In this problem, you will derive recursion relations for real-time updating of beliefs based on incoming evidence. These relations are useful for situated agents that must monitor their environments in real-time.

(a) Consider the discrete hidden Markov model (HMM) with hidden states S_t, observations O_t, transition matrix a_{ij} and emission matrix b_{ik}. Let

$$q_{it} \;=\; P(S_t\!=\!i|o_1, o_2, \ldots, o_t)$$

denote the conditional probability that S_t is in the i^{th} state of the HMM based on the evidence up to and including time t. Derive the recursion relation:

$$q_{jt} = \frac{1}{Z_t} b_j(o_t) \sum_i a_{ij} q_{it-1} \quad \text{where} \quad Z_t = \sum_{ij} b_j(o_t) a_{ij} q_{it-1}.$$

Justify each step in your derivation—for example, by appealing to Bayes rule or properties of conditional independence.

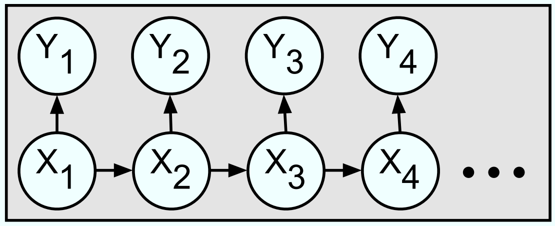
(b) Consider the dynamical system with *continuous, real-valued* hidden states X_t and observations Y_t, represented by the belief network shown below. By analogy to the previous problem (replacing sums by integrals), derive the recursion relation:

$$P(x_t|y_1, y_2, \ldots, y_t) \;=\; \frac{1}{Z_t} P(y_t|x_t) \int dx_{t-1}\, P(x_t|x_{t-1})\, P(x_{t-1}|y_1, y_2, \ldots, y_{t-1}),$$

where Z_t is the appropriate normalization factor,

$$Z_t = \int dx_t\, P(y_t|x_t) \int dx_{t-1}\, P(x_t|x_{t-1})\, P(x_{t-1}|y_1, y_2, \ldots, y_{t-1}).$$

In principle, an agent could use this recursion for real-time updating of beliefs in arbitrarily complicated continuous worlds. In practice, why is this difficult for all but Gaussian random variables?



6.4 Continuous density HMM

In class, we studied discrete HMMs with discrete hidden states and observations, as well as linear dynamical systems with continuous hidden states and observations.

This problem considers a *continuous density* HMM, which has discrete hidden states but continuous observations. Let $S_t \in \{1, 2, ..., n\}$ denote the hidden state of the HMM at time t, and let $X_t \in \Re$ denote the real-valued scalar observation of the HMM at time t. The continuous density HMM makes the same Markov assumptions as the discrete HMM in class. In particular, the joint distribution over sequences $S = \{S_t\}_{t=1}^T$ and $X = \{X_t\}_{t=1}^T$ is given by:

$$P(S, X) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \prod_{t=1}^{T} P(X_t|S_t).$$

In a continuous density HMM, however, the distribution $P(X_t|S_t)$ must be parameterized since the random variable X_t is no longer discrete. Suppose that the observations are modeled as Gaussian random variables:

$$P(X_t = x|S_t = i) \;=\; \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

with state-dependent means and variances. Indicate whether each of the following distributions is Gaussian (univariate or multivariate) or a mixture of Gaussians. Also, if the distribution is a mixture of Gaussians, indicate how many mixture components it contains. The first problem has been done as an example.

(*) $P(X_1)$

The distribution $P(X_1)$ is a *mixture* of univariate Gaussians.
It contains n mixture components because it can be written as

$$P(X_1) = \sum_{i=1}^{n} P(X_1|S_1 = i) P(S_1 = i).$$

(a) $P(X_t|S_{t-1})$

(b) $P(X_t, X_{t'}|S_t, S_{t'})$

(c) $P(X_1, X_2, \ldots, X_T)$

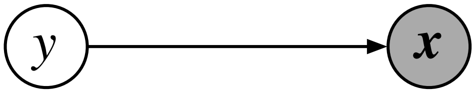
(d) $P(X_t|X_1, X_2, \ldots, X_{t-1})$

(e) $P(X_T)$

(f) $P(X_1, X_2, \ldots, X_t|S_1, S_2, \ldots, S_t)$

6.5 Mixture model decision boundary

Consider a multivariate Gaussian mixture model with two mixture components. The model has a hidden binary variable $y \in \{0, 1\}$ and an observed vector variable $\vec{x} \in \mathcal{R}^d$, with graphical model:



$$P(y\!=\!i) \;=\; \pi_i \qquad\qquad P(\vec{x}|y\!=\!i) \;=\; (2\pi)^{-\frac{d}{2}}|\Sigma_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)}$$

The parameters of the Gaussian mixture model are the prior probabilities π_0 and π_1, the mean vectors $\vec{\mu}_0$ and $\vec{\mu}_1$, and the covariance matrices Σ_0 and Σ_1.

(a) Compute the posterior distribution $P(y\!=\!1|\vec{x})$ as a function of the parameters $(\pi_0, \pi_1, \vec{\mu}_0, \vec{\mu}_1, \Sigma_0, \Sigma_1)$ of the Gaussian mixture model.

(b) Consider the special case of this model where the two mixture components share *the same* covariance matrix: namely, $\Sigma_0 = \Sigma_1 = \Sigma$. In this case, show that your answer from part (a) can be written as:

$$P(y = 1|\vec{x}) = \sigma(\vec{w} \cdot \vec{x} + b) \quad \text{where} \quad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

As part of your answer, you should express the parameters $(\vec{w}, b)$ of the sigmoid function explicitly in terms of the parameters $(\pi_0, \pi_1, \vec{\mu}_0, \vec{\mu}_1, \Sigma)$ of the Gaussian mixture model.

(c) Assume again that $\Sigma_0 = \Sigma_1 = \Sigma$. Note that in this case, the decision boundary for the mixture model reduces to a hyperplane; namely, we have $P(y = 1|\vec{x}) = P(y = 0|\vec{x})$ when $\vec{w} \cdot \vec{x} + b = 0$. Let $k > 0$. Show that the set of points for which

$$\frac{P(y = 1|\vec{x})}{P(y = 0|\vec{x})} = k$$

is also described by a hyperplane, and find the equation for this hyperplane. (These are the points for which one class is precisely k times more likely than the other.) Of course, your answer should recover the hyperplane decision boundary $\vec{w} \cdot \vec{x} + b = 0$ when $k = 1$.