Handout 6

STAT5710 Applied Stochastic Processes

Here we describe latent modeling using Markov chains. We cover the basics of hidden Markov models.

The model

A **hidden Markov model** is a process $\{X_t\}$ in which each observation is drawn from a distribution determined by the current state of an unobserved Markov chain $\{Y_t\}$.

1. Fill in the two components.
   (a) Given the states, the observations are conditionally independent, with
   $$X_t \mid Y_t = j \ \sim$$
   where f_j is an emission distribution, discrete or continuous.
   (b) $\{Y_t\}$ is _____ on a finite state space $S = \{1, \dots, m\}$.
   (c) Sketch the dependence structure, with the states along the bottom and the emissions above.

2. $\{X_t\}$ is not a Markov chain, but the pair $Z_t = (X_t, Y_t)$ is. Write its transition density.
   $$p(z_t \mid z_{t-1}) \ = \ p(z_t \mid y_{t-1}) \ =$$

Let $\nu = (\nu_1, \dots, \nu_m)^T$ denote the initial distribution of the chain, so that the model parameters are $\Theta = (\nu, P, \theta)$.

3. Write the joint density of a sample path of pairs.
   $$f((x_0, y_0), \dots, (x_T, y_T) \mid \Theta) \ =$$

The likelihood

4. Marginalize over the latent paths, then collect the sums by time and write the result as a product of matrices.

$$L(\Theta; x_0, \ldots, x_T) \;=\; \sum_{y_0=1}^{m} \cdots \sum_{y_T=1}^{m}$$

$$=$$

$$=$$

where $F(x) = \text{diag}($ $)$

5. Reading that product from the left gives the **forward recursion**.
   (a) $\alpha_t \;=$
   (b) So the likelihood is $L(\Theta; x_0, \ldots, x_T) =$ _____, which we can evaluate for any Θ and maximize numerically.

Inferring the states

The MLE does not identify which state generated a particular observation. To find states, we need to compute the posterior probability of Y_t.

6. Split the record at t, into what happened up to t and what happened after.

$$P(Y_t = j,\ x_0, \dots, x_T) \ = \ \underbrace{\qquad\qquad}_{\beta_t(j)} \cdot \underbrace{\qquad\qquad}_{\alpha_t(j)}$$

   (a) The second factor is the forward recursion from item 5. The first is given by the **backward recursion**,

$$\beta_t \ =$$

7. **Filtering** gives state probabilities conditional on the path up to t.

$$\gamma_t(j) \ := \ P(Y_t = j \mid x_0, \dots, x_t) \ = \ \frac{P(Y_t = j,\ x_0, \dots, x_t)}{P(x_0, \dots, x_t)} \ =$$

8. **Smoothing** gives state probabilities conditional on the whole path.

$$P(Y_t = j \mid x_0, \dots, x_T) \ =$$

9. So pointwise **decoding** gives $\hat{Y}_t = \arg\max_j$

   This minimizes expected misclassification, but may produce a path that is itself unlikely because…

10. **Viterbi decoding** instead maximizes $P(y_0, \ldots, y_T \mid x_0, \ldots, x_T)$ jointly over paths $y_0, \ldots, y_T$.

    (a) Define

    $$\delta_0(j) \; = \qquad\qquad \delta_t(j) \; =$$

    (b) Then start at the end and work backwards

    $$\hat{Y}_T =$$

    $$\hat{Y}_t =$$