CSC 446 Notes: Lecture 17

1 Hidden Markov Models

A Hidden Markov Model (HMM) is a Markov Chain (a series of states with probabilities of transitioning from one state to another) where the states are hidden (latent) and each state has an emission as a random variable. The model is described as follows:

- Ω : the set of states, with $y_i \in \Omega$ denoting a particular state
- Σ : the set of possible emissions with $x_i \in \Sigma$ denoting a particular emission
- $P \in \mathbb{R}_{[0,1]}^{\Omega \times \Omega}$: the matrix with each element giving the probability of a transition
- $Q \in \mathbb{R}_{[0,1]}^{\Omega \times \Sigma}$: the matrix with each element giving the probability of an emission
- Π : the matrix with each element giving the probability of starting in each state

The probability distribution of an HMM can be decomposed as follows:

$$P(x_1, \dots x_n, y_1, \dots, y_n) = \Pi(y_1) \prod_{i=1}^{n-1} P(y_i, y_{i+1}) \prod_{i=1}^{n} Q(y_i, x_i)$$

An example HMM is given:

$$\Omega = \{1, 2\}$$

$$\Sigma = \{a, b, c\}$$

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}$$

One possible sequence of observations would be:

$$1\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 2$$

$$\downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow\ \downarrow$$

$$a\ b\ c\ a\ a\ a\ a\ a\ b$$

We can consider multiple problems relating to HMMs.

1. Decoding I: Given $x_1, \dots, x_n, P, Q, \Pi$, determine the sequence $y_1 \dots y_n$ that maximizes $P(Y_1, \dots Y_n | X_1, \dots X_n)$.
2. Decoding II: Given $x_1, \dots, x_n$ and t, determine the distribution of y_K, that is, for all values a of y_t, $P(y_t = a | X_1, \dots, X_n)$.
3. Evaluation: Given $x_1, \dots x_n$, determine $P(X_1, \dots X_n)$.
4. Learning: Given a sequence of observations, $x_1^{(1)}, \dots x_n^{(1)}, \dots x_1^{(k)}, \dots x_n^{(k)}$, learn P, Q, Π that maximize the likelihood of the observed data.

We define two functions, α and β.

$$\alpha^t(a) := P(X_1 = \hat{x}_1, \ldots, X_t = \hat{x}_t, Y_t = a)$$

$$\beta^t(a) := P(X_{t+1} = \hat{x}_{t+1}, \ldots, X_n = \hat{x}_n \mid Y_t = a)$$

which are also recursively defined as follows:

$$\alpha^{t+1}(a) = \sum_{c \in \Omega} \alpha^t(c) P(c, a) Q(a, \hat{x}_{t+1})$$

$$\beta^{t-1}(a) = \sum_{c \in \Omega} Q(c, \hat{x}_t) \beta^t(c) P(a, c)$$

We return to the Decoding II problem. Given $x_1, \ldots, x_n$ and t, determine the distribution of Y_K, that is, for all values a of Y_t, $P(Y_t = a | X_1, \ldots, X_n)$. To do this, we rewrite the equation as follows:

$$P(y_t = a | X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n, Y_t = a)}{P(X_1, \ldots, X_n)}.$$

However, we need to calculate $P(X_1, \ldots, X_n)$. We can do this using either α or β.

$$P(X_1, \ldots, X_n) = \sum_{a \in \Omega} \alpha^n(a)$$

$$= \sum_{a \in \Omega} \beta^1(a) \Pi(a) Q(a, \hat{x}_1)$$

The Decoding I problem can be solved with Dynamic Programming. (Given $x_1, \ldots, x_n, P, Q, \Pi$, determine the sequence $y_1 \ldots y_n$ that maximizes $P(Y_1, \ldots Y_n | X_1, \ldots X_n)$.) We can fill in a table with the following values:

$$T[t, a] = \max_{y_1 \ldots y_t, y_t = a} P(y_1, \ldots y_t | X_1, \ldots X_t)$$

which means that each value is the probability of the most likely sequence at time t with the last emission being a. This can be computed using earlier values with the following formula:

$$T[t+1, a] = \max_{c \in \Omega} T[t, c] P(c, a) Q(a, \hat{x}_{t+1})$$

To compute the most likely sequence, we simply solve

$$\max_{a \in \Omega} T[n, a]$$

The learning problem can be solved using EM. Given the number of internal states, and $x_1, \ldots x_n$, we want to figure out P, Q, and Π. In the E step, we want to compute an expectation over hidden variables:

$$L(\theta, q) = \sum_y q(y|x) \log \frac{P(X, Y|\theta)}{q(Y|X)}$$

For HMM's, the number of possible hidden state sequences is exponential, so we use dynamic programming to compute expected counts of individual transitions and emissions:

$$P(a, b) \propto \sum_{i=1}^{n-1} q(Y_i = a, Y_{i-1} = b | X_1 \ldots X_n) \tag{1}$$

$$Q(a, b) \propto \sum_{i=1}^{n} q(Y_i = a | X_1 \ldots X_n) I(X = w) \tag{2}$$

The new P is defined as:

$$P^{\text{new}}(a, b) \propto \sum_{i=1}^{n-1} \alpha^i(a) P^{\text{old}}(a, b) \beta^{i+1} Q^{\text{old}}(b, \hat{x}_{i+1})$$

Ian Perera 4/12; DG 4/13