CS281A/Stat241A: Statistical Learning Theory

Bedtime Stories: Introduction to Hidden Markov Models

Lecturer: Michael I. Jordan, 10/28/02 *Scribes: Jayanth and Anshuman*

In this lecture, we will discuss Hidden Markov Models, and the most natural way to introduce them will be to move from conditional mixture models, which we already know about, to HMMs.

1 Conditional Mixture Models

These models attempt to model the distribution of Y as a function of X by introducing a multi-nomial variable Z between them. The idea is that for each possible value of Z, Y is modelled by a different distribution. In other words, Z decides based on the value of X, which *regime* it belongs to, and then generates Y. This is represented in the graphical model in Figure 1 (a) . Note that Z_n is not observed. For an example of data, where this model applies, see Figure 1 (b). In Regime 1, the dependence follows a diferent pattern in the other regime. The classification function decides which function should be used to generate Y. The parameters required to represent this model are simply $\theta = (\beta_j, \sigma_j)$ for each possible value j of the multinomial variable Z_n. Thus, it is not necessary to resort to non-parametric methods like kernel functions to model such cases, one can use conditional mixture models.

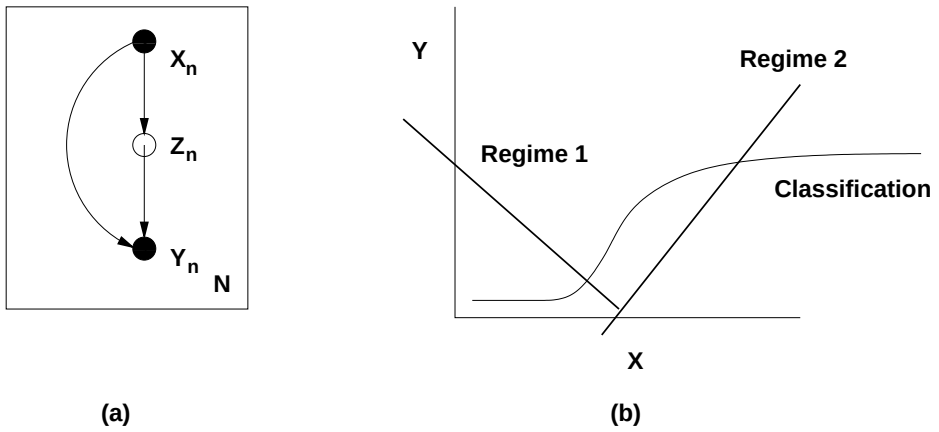

Figure 1: Conditional Mixture Models: (a) Graphical Model (b) Possible scenario of application

In the case above, with 2 regimes, the multinomial variable can take two values. The posterior function $p(Z_n \mid X_n)$ can be a logistic function or a softmax function, while any GLIM may be used for the function $P(Y_n \mid X_n, Z_n)$. In this case, the EM algorithm would first attempt to calculate the expected value of the complete log likelihood $< log[p(Y, Z \mid X, \theta] >$. Since Z is an indicator value, $E[Z_n] = P[Z_n = 1]$. The E step would attempt to calculate $p(Z_n \mid X_n, Y_n, \theta_t) = \tau$, and then the problem of estimating the parameters for $P(Y_n \mid X_n, Z_n)$ will be a weighted linear regression problem (for example).

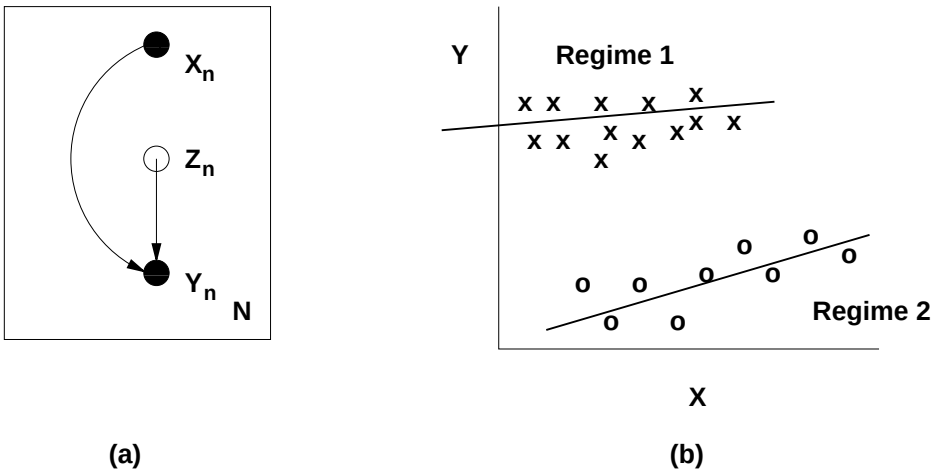

Figure 2: Switching Model: (a) Graphical Model (b) Possible scenario of application

1.1 Related Models

There are other models which are closely related to the conditional mixture model. One of them is the switching model. The graphical model and an example scenario are shown in Figure 2. This model, typically used in economics, differs from the conditional mixture model in that, Z_n is independent of X_n (though it is conditionally dependent on X_n given Y_n, as is clear from the graph). Some kind of clustering algorithm can be used to discover the clusters in the $X - Y$ space, which can be fitted using some regression model. Each of these clusters would correspond to a different value of Z_n.

Another closely related model is the decision tree model. This decision tree model differs from the conventional decision trees in that soft splitting is used conditional on the value of X_n and the decisions made earlier. Each node tests a predicate and finally the leaves may even have some constant value. This model is also called the hierarchial mixtures of experts. An sample graphical model is shown in Figure 3. The first decision $Z_n^{(1)}$ is made conditional on X_n and this can be modelled by using a logistic function (and regression used to obtain these parameters). The decisions downstream are made using the previous decisions and these will have to be modelled by other regression models.

2 Hidden Markov Models

Hidden Markov models will be our first example of non IID data sets, and it is a extremely powerful model used in a variety of applications, like part of speech tagging, speech recognition. Hidden Markov models can be thought of as a dynamic version of a switching model where the multinomial variable Z_n probabilistically changes with time, dependent only on its previous value. That is, the state variable Z_n follows a Markov chain, whose transition matrix is part of the parameters we will have to infer. It is referred to as a hidden Markov model because the state variable Z_n is not observable (to us mortals anyway). The parameters for this model is $\theta = \pi, A, \mu_i, \sum_i$, were π is the distribution of the state of time 0, $A = a_{i,j} = P(Z_{n+1} = j \mid Z_n = i)$ is the transition probability matrix, $\mu_i, \sum_i$ are the means and covariances for each value of the state i, for the probability $p(X_n|Z_n = i)$.

The graphical model of representing a hidden Markov model is shown in Figure 4 (a). The top part of this

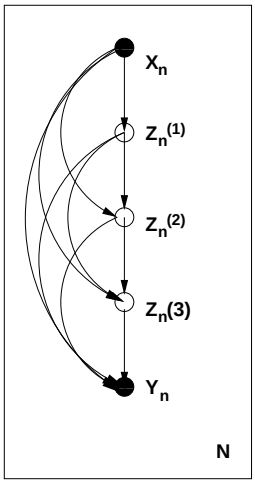

Figure 3: Decision Tree: Graphical Model

figure is a simple Markov chain and the edges from Z_n to X_n describe a generativ model. There are a number of ways of representing a hidden Markov model apart from the graphical model we are familiar with it. One is the stochastic automation shown in Figure 4 (b). This is the finite automaton garnished with probabilities for every possible transition instead of a input symbol. Another would be the lattice diagram as shown in Figure 4 (c), which expresses the state Z_n as a vector of $Z_n^{(i)}$ where $Z_n^{(i)}$ is 1 if the state is i at time n.

We will discuss the EM algorithm for HMMs in the next class. It is interesting to note that the sum product algorithm which is used in the E step (for probabilistic inference) had already been discovered for HMMs. Indeed, we will see the relation between their algorithms and the more recent sum-product algorithms.

3 Expectation Maximization Algorithm

In this lecture, we will introduce the EM algorithm as a coordinate ascent algorithm on an auxilary function, which is, as we will show, the lower bound for the log likelihood.

3.1 Our Setup

We have X denoting the observed varibales and Z indicating the latent variables. For a moment, assume that we can observe Z as well. The ML estimation problem then takes the following form:

$$l_c(\theta; x, z) = \log p(x, z|\theta), \tag{1}$$

The expression above is known as the *complete log likelihood*. However, given that Z is not observed, the complete log likelihood is indeed a random number and we cannot maximize it. To get around this we average over z to get the *expected complete log likelihood*:

$$\langle l_c(\theta; x, z)\rangle_q = \sum_z q(z|x, \theta)\log p(x, z|\theta), \tag{2}$$

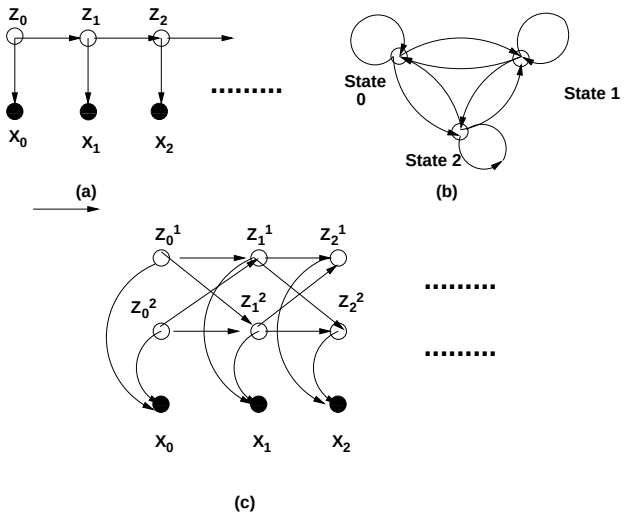

Figure 4: Hidden Markov Models (a) Graphical Model (b) Stochastic Automation (c) Lattice Diagram

The main idea behind the EM algorithm is to iteratively maximize this quantity which in essence should also maximize the real likelihood. This process is called coordinate ascent and we will show this now.

Theorem 1 *KL divergence* ≥ 0, *and equal to 0 iff* $p \equiv q$

Proof:

$$0 = log1 = log \sum_x q(x) \tag{3}$$

Then by Jensen's inequality

$$\sum_x p(x) log \frac{p(x)}{q(x)} \geq 0 \tag{4}$$

iff $p \equiv q$, then $KL = 0$.

Now we use this handy result to show that the EM algorithm is indeed an example of coordinate ascent.

observed variables: X
latent variables: Z
parameters: θ
probability model: $p(x, z|\theta)$

E step: compute $\langle l_c(\theta|x, z)\rangle$ using Equation (2)
M step: maxmimize $\langle l_c(\theta|x, z)\rangle$ w.r.t θ

We define $\mathcal{L}(q,\theta)$ as a lower bound on the true likelihood $l(\theta|x)$ for any q

$$\begin{aligned} l(\theta|x) &= \log p(x|\theta) && (5)\\ &= \log \sum_z p(x,z|\theta) && (6)\\ &= \log \sum_z q(z|x)\frac{p(x,z|\theta)}{q(z|x)} && (7)\\ &\geq \sum_z q(z|x)\log\frac{p(x,z|\theta)}{q(z|x)} && (8)\\ &= \mathcal{L}(q,\theta) && (9) \end{aligned}$$

We define $\mathcal{D}$:

$$\mathcal{D} = \sum_z q(z|x)\log p(x,z|\theta) - \sum_z q(z|x)\log q(z|x) \tag{10}$$

The first term on the right above should be familiar. It is indeed the *expected complete log likelihood* the second term does not depend on θ and hence to maximize $\mathcal{L}(q,\theta)$ we can mazimize the $\langle l_c(\theta|x,z)\rangle$ w.r.t θ. This forms our M-step

For the E-step, we need to show that the difference between $l(\theta|x)$ and $\mathcal{L}(q,\theta)$ is a KL divergence. The steps follow:

$$\begin{aligned} l(\theta|x) - \mathcal{L}(q,\theta) &= \textstyle\sum_z q(z|x)\log p(x|\theta) - \sum_z q(z|x)\log\frac{p(x,z|\theta)}{q(z|x)} && (11)\\ &= \textstyle\sum_z q(z|x)\log\frac{p(x|\theta)q(z|x)}{p(x,z|\theta)} && (12)\\ &= \textstyle\sum_z q(z|x)\log\frac{q(z|x)}{p(z|x,\theta)} && (13)\\ &= (D)(q(z|x)||p(z|x,\theta) && (14) \end{aligned}$$

Thus if we pick a distribution q, such that $q(z|x) = p(z|x,\theta^{(t)})$
we will have $l(\theta|x) = \mathcal{L}(q,\theta^{(t)})$

The EM algorithm is thus a coordinate ascent in $\mathcal{L}(q,\theta)$, with the two steps being:

E-step:

$$q^{(t+1)}(z|x) = p(z|x,\theta^{(t)}) = argmax_q \mathcal{L}(q,\theta^{(t)}) \tag{15}$$

M-step:

$$\theta^{(t+1)} = argmax_\theta \mathcal{L}(q^{(t+1)},\theta) \tag{16}$$