Joint and Conditional probability models

Ananth Ranganathan

July 5th 2006

There are two main ways to learn a probabilistic model of some stochastic process given some data about the input/output characteristics of the process (what I call input and output can also be called the measurement and state of the system). If the aim is to understand how the various components of the input and output interact and answer all possible questions about such interactions, the appropriate model to build is a generative one. On the other hand, if the purpose of the modeling is to answer a specific question or solve a specific problem (for example, what is the probability of a particular output given some input?) then the more apt model is a discriminative one.

While purely discriminative methods simply learn a model that optimizes the input/output relationship on a given training set, this can be done in a probabilistic context through the use of Conditional models instead of complete generative models. Here we summarize the differences between generative and conditional models.

In the following discussion, I will represent the measurements by z and the state by x. We also have available a set of training data $\{X, Z\}$ where $Z = \{z_1, \dots, z_n\}$ are measurements, and the corresponding states are $X = \{x_1, \dots, x_n\}$.

The generative model for a stochastic process as described above is the joint (posterior) distribution $p(x, z|X, Z)$. This distribution gives us a complete description of the stochastic process. A predictive model on the output (states) is then given by

$$\begin{aligned}
p(x|z, X, Z) &= \frac{p(x, z|X, Z)}{p(z|X, Z)} \\
&= \frac{p(x, z|X, Z)}{\int_x p(x, z|X, Z)} \\
&= \frac{\int_{\theta_j} p(x, z, \theta_j|X, Z)}{\int_{x,\theta_j} p(z, x, \theta_j|X, Z)} \\
&= \frac{\int_{\theta_j} p(x, z|\theta_j, X, Z)p(\theta_j|X, Z)}{\int_{x,\theta_j} p(z, x|\theta_j, X, Z)p(\theta_j|X, Z)} \\
&= \frac{\int_{\theta_j} p(x|\theta_j, z, X, Z)p(z|\theta_j, X, Z)p(\theta_j|X, Z)}{\int_{\theta_j} p(z|\theta_j, X, Z)p(\theta_j|X, Z)} \qquad (1)
\end{aligned}$$

where it is assumed that the joint distribution is parametrized by θ_j and by integrating over θ_j we essentially take into account all possible functional forms of the joint distribution $p(x, z|X, Z)$.

In most cases, the integrations in (1) cannot be performed easily, and hence we resort to learning the model through either the Maximum Likelihood (ML) or the Maximum a Posteriori (MAP) paradigms.

$$\begin{aligned} p(x|z, X, Z) &= p(x|\theta^*, z, X, Z) \\ \text{where} \quad \theta^* &= \underset{\theta_j}{\operatorname{argmax}} \; p(z|\theta_j, X, Z)p(\theta_j|X, Z) \qquad \text{(MAP)} \\ \text{and} \quad \theta^* &= \underset{\theta_j}{\operatorname{argmax}} \; p(z|\theta_j, X, Z) \qquad \text{(ML)} \end{aligned} \tag{2}$$

These computations can be performed in closed form for a few families of distributions, the most famous being the exponential family, which gives rise to a convex optimization problem. In this case, the ML result can be shown to be the same as that obtained using the Maximum Entropy principle. In other cases where closed form solution is not possible, iterative techniques such as EM and gradient descent are employed.

The other way to obtain the conditional density is to model the conditional density directly without involving the joint as follows -

$$\begin{aligned} p(x|z, X, Z) &= \int_{\theta_c} p(x, \theta_c|z, X, Z) \\ &= \int_{\theta_c} p(x|z, \theta_c, X, Z)p(\theta_c|X, Z) \end{aligned} \tag{3}$$

whence we can obtain ML and MAP solutions similarly. The difference between (3) above and the generative model (1) can be seen more clearly using the following decomposition of the prior on θ_c, $p_c(\theta|X, Z)$

$$\begin{aligned} p(\theta_c|X, Z) &= \frac{p(X|\theta_c, Z)p(\theta_c|Z)}{p(X|Z)} \\ &= \frac{p(X|\theta_c, Z)p(\theta_c, Z)}{p(X, Z)} \\ &= \frac{p(X|\theta_c, Z)p(\theta_c)p(Z)}{p(X, Z)} \end{aligned}$$

Note the crucial difference on the last line where $p(\theta_c, Z) = p(\theta_c)p(Z)$. This is a direct result of the model being conditional so that θ_c is independent of Z if X is unknown. This independence does not hold for the generative model. The difference in graphical models is illustrated in Figure 1.

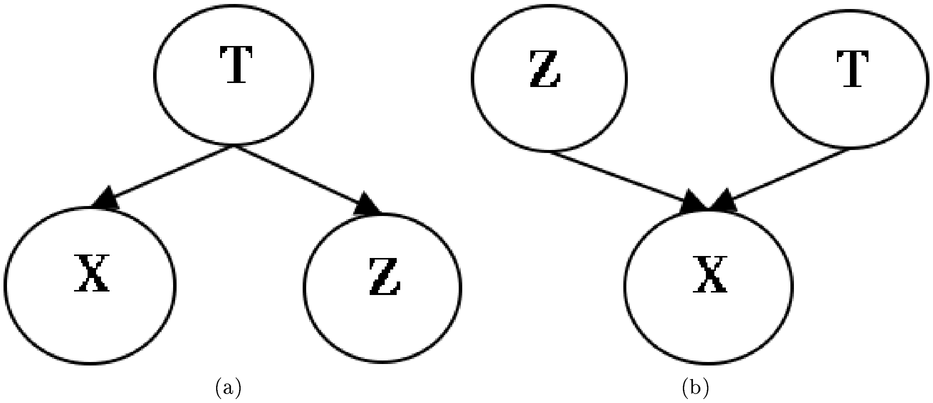

Figure 1: The graphical models that represent the case for the (a) joint distribution and (b) conditional distribution. (T represents θ)