CS281B/Stat241B: Advanced Topics in Learning & Decision Making

More on Marginal Likelihood

Lecturer: Michael I. Jordan *Scribes: Shawn Chang and Barry Chen*

Today we will be taking a closer look at the marginal likelihood.

$$p(\mathcal{D}|M) = \int d\theta \; p(\mathcal{D}, \theta|M) \tag{1}$$

In particular, we will see another interpretation of the marginal likelihood that will give more intuition, and we will learn how to calculate the marginal likelihood for a fully-observed, discrete graphical model.

1 Prequential Interpretation of the Marginal Likelihood

In the prequential interpretation of the marginal likelihood, we view the marginal likelihood as a measure of how well a given model predicts future data given a sequence of data points. This interpretation naturally arises from using the chain rule of probability to express the marginal likelihood. Recall that $\mathcal{D} \triangleq (x_1, x_1, \ldots, x_N)$. Using the chain rule of probability and the multiplication rule for log, we can express the log marginal likelihood as:

$$\begin{aligned} \log p(\mathcal{D}|M) &= \log \left[p(x_1|M) \prod_{n=2}^{N} p(x_n|x_1, \ldots, x_{n-1}, M) \right] & (2) \\ &= \log p(x_1|M) + \sum_{n=2}^{N} \log p(x_n|x_1, \ldots, x_{n-1}, M) & (3) \end{aligned}$$

Intuitively, $p(x_n|x_1, \ldots, x_{n-1}, M)$ is a measure of how good your model M does at predicting the n_{th} data point based all the previous data points. Given any ordering of your data $\mathcal{D}$, you can think of $\log p(x_1|M) + \sum_{n=2}^{N} \log p(x_n|x_1, \ldots, x_{n-1}, M)$ as the negative sum of log loss. The smaller this sum is, the worse M is as a model for this data.

Notice that this interpretation has a similar flavor to cross-validation. Like cross-validation, the marginal likelihood calculates a measure of how well the model predicts the data. In cross-validation we measure the total loss over all the cross-validation folds, and from the prequential interpretation of the marginal likelihood, we sum the loss arising from the prediction of a new data point based on all the past data points. Take note, however, that cross-validation and the prequential interperetation differ in the following way. Typical n-fold cross-validation uses a held out data set for evaluation purposes in one fold, and then uses the same set for training in the other folds, whereas, in the prequential interpretation the only data used to predict a new data point are the old data points. For example, cross-validation will calculate the loss based on $p(x_2|x_1, M)$ and $p(x_1|x_2, M)$, but in the prequential interpretation only $p(x_2|x_1, M)$ will be used in the loss calculation if for example $\mathcal{D}$ was ordered as (x_1, x_2).

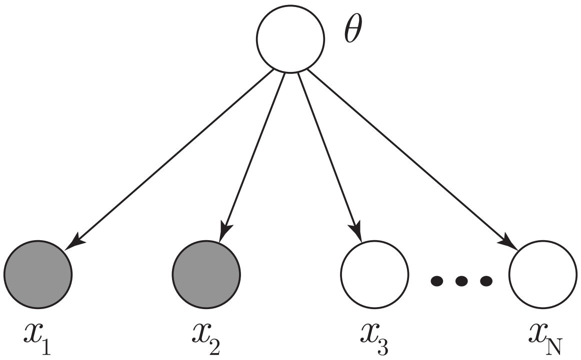

Figure 1: Calculating the Posterior on θ

1.1 Prequential Interpretation from a Bayesian Perspective

The expression, $p(x_n|x_1, \ldots, x_{n-1}, M)$ is a prediction based on marginalizing over the parameters, i.e.,

$$\begin{aligned} p(x_{new}|\chi_{old}) &= \int d\theta \; p(x_{new}|\theta, \chi_{old}) p(\theta|\chi_{old}) & (4) \\ &= \int d\theta \; p(x_{new}|\theta) p(\theta|\chi_{old}) & (5) \end{aligned}$$

where the second step comes from conditional independence of the data given the parameter θ (see figure 1 for graphical setup). At each step, as you get new data, your belief on the values of θ may change, and so a Bayesian includes the posterior $p(\theta|\chi_{old})$ in the prediction calculation. Now to do model selection, the Bayesian calculates:

$$p(M|\mathcal{D}) \propto P(\mathcal{D}|M)P(M) \tag{6}$$

and maximizes over M.

2 How Do We Calculate Marginal Likelihoods?

2.1 A Little Notation

Consider an arbitrary graphical model with the following setup (see figure 2):

- All nodes observed
- All nodes are discrete

  Nodes are distributed as a multinomial distribution and takes on k possible vectors,i.e., the i_{th} node can take on:

  $x_i \in \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\}$ Where x_i is a vector of size k.

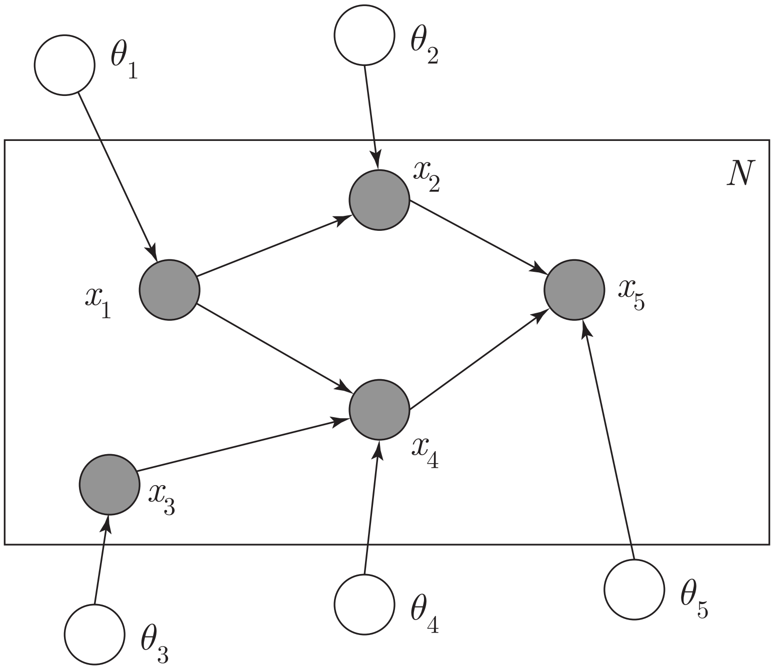

Figure 2: Calculating the Posterior on θ

The probability of a particular configuration of all the nodes x is:

$$p(x) = \prod_{i \in nodes} p(x_i | x_{\pi_i}) \tag{7}$$

- Use Conjugate Priors for the Multinomial: Dirichlet Priors

  Conjugate priors keep calculations of posteriors within the same family of distributions. For more details on conjugate priors, consult Chapter 4 of Jordan and Bishop. The Dirichlet priors are parameterized by α_{ijk}, where α_{ijk} is the Dirichlet parameter for node i in its j_{th} configuration and its parents in their k_{th} configuration.

  $\alpha_{ij} \triangleq \sum_k \alpha_{ijk}$

- Conditional Probability

  $\theta_{ijk} \triangleq$ Probability of node i being in configuration k, given its parents are in their j_{th} configurations (where j runs from 0 to $k^M - 1$ [M parents taking on k values each]).

- Sufficient Statistics: Counts $m_{ijk} \triangleq$ count of the number of times node i takes k_{th} value and parents in j_{th} configuration

  $m_{ij} \triangleq \sum_k m_{ijk}$

2.2 The Marginal Likelihood Using Dirichlet Priors

Using the above setup, the marginal likelihood for an arbitrary, discrete, fully-observed graphical model using Dirichlet priors is:

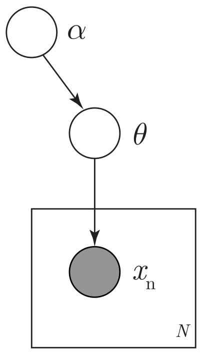

Figure 3: A Single Discrete Node Graphical Model

$$p(\mathcal{D}|M) = \prod_i \prod_j \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + m_{ij})} \prod_k \frac{\Gamma(\alpha_{ijk} + m_{ijk})}{\Gamma(\alpha_{ijk})} \tag{8}$$

where $\Gamma(n) = (n-1)!$ is the gamma function.

2.3 Derivation

To see how to derive the closed-form solution of the marginal likelihood above, let's consider a single discrete node like the one in figure 3.

Define $p(x_n^k = 1|\theta) \triangleq \theta_k$, where x_n^k is the value of the k_{th} element in the n_{th} sample of x and $\sum\limits_k \theta_k = 1$. The sufficient statistics are the counts: $m_k \triangleq \sum\limits_{n=1}^{N} x_n^k$

The likelihood of a single sample is given by:

$$p(x_n|\theta) = \theta_1^{x^1_n} \theta_2^{x^2_n} \dots \theta_k^{x^k_n} \tag{9}$$

and the total likelihood is:

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{n=1}^{N} p(x_n|\theta) && (10) \\ &= \theta_1^{m_1} \theta_2^{m_2} \dots \theta_k^{m_k} && (11) \end{aligned}$$

The prior of the parameter θ is the Dirichlet prior for this discrete node.

$$p(\theta|\alpha) = \mathrm{C}(\alpha) \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1} \tag{12}$$

$$\mathrm{Dir}(\theta|\alpha_1, \alpha_2, \dots, \alpha_k) \triangleq \mathrm{C}(\alpha) \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1} \tag{13}$$

Using integration by parts, one can derive the normalizing constant $\mathrm{C}(\alpha)$:

$$\mathrm{C}(\alpha) = \frac{\Gamma(\sum\limits_k \alpha_k)}{\prod\limits_k \Gamma(\alpha_k)} \tag{14}$$

We can now calculate $p(\theta, \mathcal{D}|\alpha)$:

$$\begin{aligned} p(\theta, \mathcal{D}|\alpha) &= p(\mathcal{D}|\theta, \alpha)p(\theta|\alpha) && (15) \\ &= p(\mathcal{D}|\theta)p(\theta|\alpha) && (16) \\ &= \mathrm{C}(\alpha)\theta_1^{\alpha_1+m_1-1}\theta_2^{\alpha_2+m_2-1}\dots\theta_k^{\alpha_k+m_k-1} && (17) \\ &= \mathrm{Dir}(\theta|\alpha_1+m_1, \alpha_2+m_2, \dots, \alpha_k+m_k) && (18) \end{aligned}$$

This leads directly to the marginal likelihood:

$$\begin{aligned} p(\mathcal{D}|\alpha) &= \int d\theta \; p(\mathcal{D}, \theta|\alpha) && (19) \\ &= \int d\theta \mathrm{C}(\alpha)\theta_1^{\alpha_1+m_1-1}\theta_2^{\alpha_2+m_2-1}\dots\theta_k^{\alpha_k+m_k-1} && (20) \\ &= \mathrm{C}(\alpha)\left(\frac{\Gamma(\sum_k \alpha_k + m_k)}{\prod_k \alpha_k + m_k}\right)^{-1} && (21) \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k + m_k)}\prod_k \frac{\Gamma(\alpha_k + m_k)}{\Gamma(\alpha_k)} && (22) \end{aligned}$$

Taking the product of this expression over the nodes in the graph (a product over i) and over the parent configurations of each node (a product over j) yields Equation 8.