CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Bayesian Model Selection

Lecturer: Michael I. Jordan *Scribes: David Blei and Jon McAuliffe*

1 Introduction

Bayesian methods treat the model M as a random variable on a set of models $\mathcal{M}$. Let $\mathcal{D}$ denote the data $\{X_1, X_2, \ldots, X_N\}$. To make predictions, the Bayesian statistician would ideally integrate over all models $M \in \mathcal{M}$:

$$P(X \mid \mathcal{D}) = \int_M \int_\theta P(X|\theta, M)P(\theta|M, \mathcal{D})P(M \mid \mathcal{D})d\theta dM$$

a procedure which is called *model averaging*. Model averaging is often prohibitively expensive computationally, however, and one often resorts to selecting a single model M^*:

$$M^* = \arg\max_M P(M|\mathcal{D})$$

on which to base predictions. The selection of a single model—*model selection*—can also be justified in a decision-theoretic treatment of the Bayesian paradigm.

2 The MAP Estimate

The maximum a posteriori (MAP) estimate of the parameters maximizes the posterior parameter distribution:

$$\hat{\theta}_{\text{MAP}} = \arg\max_\theta P(\theta|\mathcal{D}, M)$$

The posterior distribution of the parameters is

$$P(\theta|\mathcal{D}, M) = \frac{P(\mathcal{D}|\theta, M)P(\theta|M)}{P(\mathcal{D}|M)}$$

Note that this is composed of the likelihood of the data, prior probability of the parameters, and marginal likelihood of the data. As usual, we take the log of the parameter posterior.

$$\log P(\theta|\mathcal{D}, M) = \log P(\mathcal{D}|\theta, M) + \log P(\theta|M) + C$$

The constant factor C does not depend on θ. The MAP estimate of the parameter is given by

$$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \arg\max_\theta [P(\theta|\mathcal{D}, M)] \\ &= \arg\max_\theta [\log P(\mathcal{D}|\theta, M) + \log P(\theta|M)] \end{aligned}$$

Note that without the prior, $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{ML}}$.

Example *Gaussian prior* Suppose $P(\theta|M) = \exp\left[-\frac{\lambda}{2}(\theta^T\theta)\right]$

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta}\left[\log P(\mathcal{D}|\theta, M) - \frac{\lambda}{2}||\theta||_2^2\right]$$

The penalty term is exactly the L_2 regularizer.

Example *Laplace prior* Suppose that $P(\theta|M) = \exp\{-\lambda\sum|\theta_i|\}$.

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta}\{\log P(\mathcal{D}|\theta, M) - \lambda\sum|\theta_i|\}$$

This penalty term is the L_1 regularizer.

3 Model Selection

Under Bayesian model selection, we would like to find the probability model which maximizes the model posterior:

$$P(M|\mathcal{D}) = \frac{P(\mathcal{D}|M)P(M)}{P(\mathcal{D})}$$

The $P(M)$ term is a prior on models and is often assumed uniform. Furthermore, $P(\mathcal{D})$ is a constant with respect to M and can be dropped. Therefore, the goal of Bayesian model selection is to compute $P(\mathcal{D}|M)$. This quantity is given by the following integral:

$$P(\mathcal{D}|M) = \int_\theta P(\mathcal{D}|\theta, M)P(\theta|M)d\theta$$

These days, most Bayesians use sampling methods to compute this integral. We will talk about sampling methods in later lectures; for now we discuss an older, deterministic approximation technique, the Bayesian information criterion (BIC).

First, we approximate the parameter posterior locally in θ space by taking the second order Taylor expansion of the log posterior. At $\hat{\theta}_{\text{MAP}}$, the derivative of $P(\theta|\mathcal{D}, M)$ is 0 by definition. When we expand about $\hat{\theta}_{\text{MAP}}$ in the Taylor approximation, the first order term vanishes:

$$\begin{aligned}
\log P(\theta|\mathcal{D}, M) &\approx \log P(\hat{\theta}_{\text{MAP}}|\mathcal{D}, M) - \frac{1}{2}(\theta - \hat{\theta}_{\text{MAP}})^T H(\theta - \hat{\theta}_{\text{MAP}}) \\
H &:= \frac{\partial^2 \log P(\theta|\mathcal{D}, M)}{\partial\theta\partial\theta^T}
\end{aligned}$$

Taking the exponential of this expression gives the *Gaussian approximation of the posterior*:

$$P(\theta|\mathcal{D}, M) \propto P(\mathcal{D}|\hat{\theta}_{\text{MAP}}, M)P(\hat{\theta}_{\text{MAP}}|M)\exp\left\{-\frac{1}{2}(\theta - \hat{\theta}_{\text{MAP}})^T H(\theta - \hat{\theta}_{\text{MAP}})\right\} \tag{1}$$

The next step is to approximate the marginal likelihood from the parameter posterior. Observe that $P(\theta, \mathcal{D}|M) \propto P(\theta|\mathcal{D}, M)$ when M and D are fixed. Therefore, we can use (1) to approximate the marginal likelihood:

$$\begin{aligned}
P(\mathcal{D}|M) &= \int_\theta P(\theta, \mathcal{D}|M)d\theta \\
&\approx \int_\theta P(\mathcal{D}|\hat{\theta}_{\text{MAP}}, M)P(\hat{\theta}_{\text{MAP}}|M)\exp\left\{-\frac{1}{2}(\theta - \hat{\theta}_{\text{MAP}})^T H(\theta - \hat{\theta}_{\text{MAP}})\right\} \\
&= P(\mathcal{D}|\hat{\theta}_{\text{MAP}}, M)P(\hat{\theta}_{\text{MAP}}|M)\int_\theta \exp\left\{-\frac{1}{2}(\theta - \hat{\theta}_{\text{MAP}})^T H(\theta - \hat{\theta}_{\text{MAP}})\right\} \\
&= P(\mathcal{D}|\hat{\theta}_{\text{MAP}}, M)P(\hat{\theta}_{\text{MAP}}|M)(2\pi)^{\frac{k}{2}}|H|^{-\frac{1}{2}}
\end{aligned}$$

where k is the dimensionality of the parameters.

Taking logs, we obtain the *Laplace approximation of the marginal likelihood*:

$$\log P(\mathcal{D}|M) \approx \log P(\mathcal{D}|\hat{\theta}_{\text{MAP}}, M) + \log P(\hat{\theta}_{\text{MAP}}|M) + \frac{k}{2}\log 2\pi - \frac{1}{2}\log|H| \tag{2}$$

The BIC rule approximates (2) by keeping only terms which increase with N. Furthermore, it asserts that $\hat{\theta}_{\text{MAP}} \approx \hat{\theta}_{\text{ML}}$ and approximates $\log|H|$ with $k\log N$ (this is an approximation which is true to order $O(1/N)$ under weak assumptions):

$$\log P(\mathcal{D}|M) \approx \log P(\mathcal{D}|\hat{\theta}_{\text{ML}}, M) - \frac{k}{2}\log N \tag{3}$$

It is important to notice that, though similar in form to the AIC, the BIC comes from a completely different set of assumptions and maximization criteria.