Bayesian Statistics from Probabilistic Machine Learning: An Introduction

Kevin P. Murphy

April 21, 2021

1 Introduction

This document summarizes Section 4.6, "Bayesian Statistics," from Kevin P. Murphy's "Probabilistic Machine Learning: An Introduction." It covers the fundamental concepts, including Bayes' rule, conjugate priors (specifically the Beta-Binomial model), posterior inference, and posterior predictive distributions.

2 Training and Test Error

For more complex models, the test error and training error tend to be larger, but decrease as the size of the dataset (N) grows. Interestingly, the training error might initially increase with N for sufficiently flexible models. This occurs because as the dataset expands, more distinct input-output pattern combinations are observed, making the task of fitting the data more challenging. Eventually, the training set resembles the test set, leading to convergence of the error rates and reflecting the model's optimal performance.

3 Bayesian Statistics (Section 4.6)

3.1 Overview

Bayesian statistics focuses on modeling uncertainty about parameters using probability distributions, rather than relying solely on point estimates. The posterior distribution is used to represent this uncertainty.

3.2 Bayes' Rule

The posterior distribution $p(\theta|D)$ is computed using Bayes' rule:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} = \frac{p(\theta)p(D|\theta)}{\int p(\theta')p(D|\theta')d\theta'}$$

where:

- $p(\theta)$ is the prior distribution, representing prior knowledge before observing the data.
- $p(D|\theta)$ is the likelihood function, reflecting the probability of observing the data given specific parameter values.
- $p(D)$ is the marginal likelihood, computed by marginalizing over the unknown θ. It represents the average probability of the data, averaged with respect to the prior. It's constant with respect to θ.

This is analogous to Bayes' rule for COVID-19 testing, but instead of disease state, we are inferring parameters of a statistical model, conditioned on a set of observations $D = \{(x_n, y_n) : n = 1 : N\}$ (for supervised learning) or $D = \{(y_n) : n = 1 : N\}$ (for unsupervised learning).

3.3 Posterior Predictive Distribution

The posterior predictive distribution over outputs, given inputs, is obtained by marginalizing out the unknown parameters:

$$p(y|x, D) = \int p(y|x, \theta)p(\theta|D)d\theta$$

This is a form of Bayes Model Averaging (BMA), making predictions using an infinite set of models (parameter values), weighted by their likelihood. BMA reduces the risk of overfitting.

3.4 Conjugate Priors (Section 4.6.1)

A prior $p(\theta) \in \mathcal{F}$ is a conjugate prior for a likelihood function $p(D|\theta)$ if the posterior $p(\theta|D)$ is in the same parameterized family $\mathcal{F}$. In other words, $\mathcal{F}$ is closed under Bayesian updating. Computations can be performed in closed form when $\mathcal{F}$ corresponds to the exponential family.

3.5 The Beta-Binomial Model (Section 4.6.2)

Consider tossing a coin N times and wanting to infer the probability of heads. Let $y_n = 1$ denote heads, $y_n = 0$ denote tails, and $D = \{y_n : n = 1 : N\}$ be the data. Assume $y_n \sim \text{Ber}(\theta)$, where $\theta \in [0, 1]$ is the probability of heads.

3.5.1 Bernoulli Likelihood (Section 4.6.2.1)

Assuming the data are i.i.d., the likelihood is:

$$p(D|\theta) = \prod_{n=1}^{N} \theta^{y_n}(1-\theta)^{1-y_n} = \theta^{N_1}(1-\theta)^{N_0}$$

where $N_1 = \sum_{n=1}^{N} I(y_n = 1)$ (number of heads) and $N_0 = \sum_{n=1}^{N} I(y_n = 0)$ (number of tails) are sufficient statistics. $N = N_0 + N_1$ is the sample size.

3.5.2 Binomial Likelihood (Section 4.6.2.2)

Consider a Binomial likelihood model where we observe the number of heads, y, in N trials:

$$p(D|\theta) = \text{Bin}(y|N, \theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y}$$

The inferences about θ are the same as with the Bernoulli likelihood.

3.5.3 Prior (Section 4.6.2.3)

To simplify computations, use a conjugate prior for the likelihood function. For the Bernoulli/Binomial likelihood, use a Beta distribution:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} = \text{Beta}(\theta|\alpha, \beta)$$

3.5.4 Posterior (Section 4.6.2.4)

Multiplying the Bernoulli likelihood with the Beta prior results in a Beta posterior:

$$p(\theta|D) \propto \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$$

The Beta distribution is a conjugate prior for the Bernoulli likelihood. α and β are hyperparameters (pseudo counts). The strength of the prior is controlled by $\tilde{N} = \alpha + \beta$ (equivalent sample size).

3.5.5 Example (Section 4.6.2.5)

- Setting $\alpha = \beta = 2$: Weak preference for $\theta = 0.5$. The posterior is a "compromise" between the prior and likelihood (see Figure 4.10a).
- Setting $\alpha = \beta = 1$: Uniform prior $p(\theta) = \text{Beta}(\theta|1,1) \propto \text{Unif}(\theta|0,1)$. The posterior has the same shape as the likelihood (see Figure 4.10b).

3.5.6 Posterior Mode (MAP Estimate) (Section 4.6.2.6)

The most probable value of the parameter is the MAP estimate:

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} \log p(\theta) + \log p(D|\theta)$$

$$\hat{\theta}_{\text{map}} = \frac{\alpha + N_1 - 1}{\alpha + N_1 - 1 + \beta + N_0 - 1}$$

- Beta(2,2) prior: $\hat{\theta}_{\text{map}} = \frac{N_1+1}{N+2}$ (add-one smoothing).
- Uniform prior: The MAP estimate becomes the MLE: $\hat{\theta}_{\text{mle}} = \arg\max_\theta \log p(D|\theta) = \frac{N_1}{N}$.

3.5.7 Posterior Mean (Section 4.6.2.7)

The posterior mean is a more robust estimate than the posterior mode:

$$E[\theta|D] = \frac{\alpha + N_1}{\alpha + N_1 + \beta + N_0} = \frac{\alpha + N_1}{\tilde{N} + N}$$

The posterior mean is a convex combination of the prior mean, $m = \frac{\alpha}{\tilde{N}}$, and the MLE: $\hat{\theta}_{\text{mle}} = \frac{N_1}{N}$:

$$E[\theta|D] = \frac{\tilde{N}}{\tilde{N} + N} m + \frac{N}{\tilde{N} + N} \frac{N_1}{N} = \lambda m + (1 - \lambda)\hat{\theta}_{\text{mle}}$$

where $\lambda = \frac{\tilde{N}}{\tilde{N}+N}$ is the ratio of the prior to posterior equivalent sample size.

3.5.8 Posterior Variance (Section 4.6.2.8)

The posterior variance captures the uncertainty in the estimate. The standard error is $se(\theta) = \sqrt{V[\theta|D]}$.

$$V[\theta|D] = \frac{(\alpha + N_1)(\beta + N_0)}{(\alpha + N_1 + \beta + N_0)^2(\alpha + N_1 + \beta + N_0 + 1)}$$

If $N \gg \alpha + \beta$, then $V[\theta|D] \approx \frac{\hat{\theta}(1-\hat{\theta})}{N}$. The uncertainty decreases at a rate of $1/\sqrt{N}$.

3.5.9 Posterior Predictive (Section 4.6.2.9)

A plug-in approximation (estimating parameters from training data and plugging them back into the model) can lead to overfitting. The Bayesian solution involves marginalizing out θ.

Bernoulli Model:

$$p(y = 1|D) = \int_0^1 p(y = 1|\theta)p(\theta|D)d\theta = E[\theta|D] = \frac{\alpha + N_1}{\alpha + N_1 + \beta + N_0}$$

With a uniform prior, $p(y = 1|D) = \frac{N_1+1}{N_1+N_0+2}$, which is Laplace's rule of succession.

Binomial Model:

The posterior predictive distribution is:

$$p(y|D, M) = \int_0^1 \text{Bin}(y|M, \theta)\text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta = \text{BetaBinomial}(y|M, \alpha + N_1, \beta + N_0)$$

$$\text{Bb}(x|M, \alpha, \beta) = \binom{M}{x} \frac{B(x + \alpha, M - x + \beta)}{B(\alpha, \beta)}$$

where $B(\alpha, \beta)$ is the Beta function. Bayesian predictions are less prone to overfitting.

3.5.10 Marginal Likelihood (Section 4.6.2.10)

The marginal likelihood or evidence for a model M is:

$$p(D|M) = \int p(\theta|M)p(D|\theta, M)d\theta$$

While constant with respect to θ for parameter inference, it plays a vital role when choosing between different models (as discussed in Section 5.4.2). It is also useful for estimating hyperparameters.