

---

## CS 181 LECTURE 2/13

---

Scribe Notes

# Contents

|          |                                                                |          |
|----------|----------------------------------------------------------------|----------|
| <b>1</b> | <b>Pictures of the Board</b>                                   | <b>3</b> |
| <b>2</b> | <b>Introduction Bayesian Model Selection</b>                   | <b>4</b> |
| 2.1      | Story from the World . . . . .                                 | 4        |
| <b>3</b> | <b>Exploration of Bayesian Methods</b>                         | <b>4</b> |
| 3.0.1    | Differentiating between Frequentist and Bayesian . . . . .     | 5        |
| 3.0.2    | Bayesian View . . . . .                                        | 5        |
| <b>4</b> | <b>Compute the Posterior</b>                                   | <b>5</b> |
| <b>5</b> | <b>Posterior Predictive</b>                                    | <b>7</b> |
| <b>6</b> | <b>Marginal Likelihood</b>                                     | <b>7</b> |
| 6.1      | Bayesian Occam's Razor aka Likelihood Regularization . . . . . | 7        |
| <b>7</b> | <b>Concept Check</b>                                           | <b>8</b> |

# 1 Pictures of the Board

CS181: Bayesian Model Selection  
 bit.ly/cs181-quiet-questions  
 bit.ly/cs181-attendance; code: subjective  
 HW2 is out  
 Book: 2, 8, 2.9

|                 | Supervised | prob | prob |
|-----------------|------------|------|------|
| discrete y      | ✓          | ✓    |      |
| continuous y    | ✓          | ✓    |      |
| Model Select: w | ✓          |      |      |

|                 | Frequentist | Bayes          |
|-----------------|-------------|----------------|
| Param w, $\phi$ | Fixed       | Random         |
| Data            | Random      | Fixed/Observed |

Bayesian View tells us how to:

1.  $p(w|X, Y)$  posterior over w
2.  $p(y^*|x^*, w) = \int p(y^*|x^*, w) p(w|X, Y)$
3.  $p(Y|X) = \int p(Y|X, w) p(w)$

Example: Only  $Y$ 's: coin flips  
 •  $Y=0$  or 1 (heads/tails)  
 • param:  $\theta$ =prob of heads  
 $p(y|\theta) = \theta^y (1-\theta)^{1-y}$

① Computing Posterior ✓  

$$p(w|X, Y) = \frac{p(Y|X, w) p(w)}{\int p(Y|X, w) p(w)} \propto p(Y|X, w) p(w)$$

Example, cont':  
 $P(Y|\theta) = \prod_n \theta^{y_n} (1-\theta)^{1-y_n} = \theta^{N_1} (1-\theta)^{N_0}$   
 recall:  $\theta_{MLE} = N_1 / (N_1 + N_0)$   
 define:  $p(\theta) \rightarrow$  comp. convenience Beta dist

$p(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$ ,  $\alpha, \beta > 0$

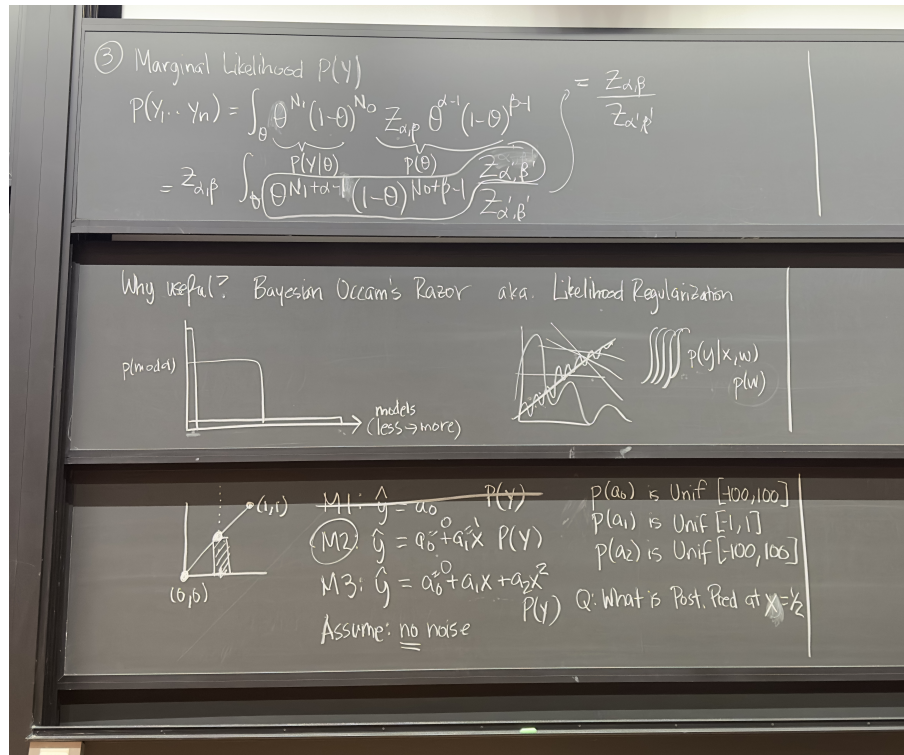
if we choose  $p(\theta)$  is Beta dist:  
 $p(\theta) p(\theta) \propto \theta^{N_1+\alpha-1} (1-\theta)^{N_0+\beta-1}$   
 New Beta!!  
 $\alpha' = N_1 + \alpha$   
 $\beta' = N_0 + \beta$

Apply:  $\alpha = \beta = 1$   
 $N_1 = 2$   
 $N_0 = 0$   
 $\alpha' = 3$   
 $\beta' = 1$

Last time: Max A Posteriori Solution:  $\arg \max_{\theta} p(Y|\theta) p(\theta)$   
 $\hookrightarrow \arg \max_{\theta} \log p(Y|\theta) + \log p(\theta)$  looks like a regularizer

In this case:  $\theta_{MAP} = \frac{N_1 + \alpha - 1}{N_1 + N_0 + \alpha + \beta - 2} = 1$   
 $\theta_{MLE} = 1$

(Apologies the second picture is a bit cut off at the bottom.)



## 2 Introduction Bayesian Model Selection

We looked at a number of basis functions in homework 1. This presents the problem of choosing a model. The natural question after seeing all the choices is how to pick the choice.

Let's start with an example.

### 2.1 Story from the World

Predict who is going to be at highest risk of complications out of patients at the ER.

Input: Patient info Output: Complications from flu

So, to create this model, we can search over different hyperparameters and choose the very best predictor. As researchers did this analysis, they found that patients with asthma were classified as low risk. But this is probably that case because the hospital will take a person's case much more seriously if they have asthma. Therefore, people with asthma end up having less complications because their treatment at the ER is more intense. Hence, this model had a missing piece of information.

## 3 Exploration of Bayesian Methods

Last lecture contains more practical information. This lecture is about expanding our minds with elegant math.

### 3.0.1 Differentiating between Frequentist and Bayesian

#### Frequentist View

There is an assumption that the parameters  $w, \phi$  are fixed. Fixed means that there is a truth, whether we know it or not.

There is an assumption that data is random. We even averaged over possible datasets.

#### Bayesian View

The Bayesian view flips this around.

The data is seen as fixed/observed. We look at the real-world.

The parameters are considered random. The idea is that there is a distribution over the parameters. There is a matter of faith here, because you specify what you think the distribution of the parameter is. You always start with your opinion and revise as you see more data.

### 3.0.2 Bayesian View

Bayesian view tells us how to:

1. Update the distribution  $p(w \mid x, y)$  posterior over  $w$
2.  $p(y^* \mid x^*, X, Y) = \int_w p(y^* \mid x^*, w)p(w \mid X, Y)$  The integral is averaging over the possible  $w$ .
3.  $p(Y \mid X) = \int_w p(Y \mid X, w)p(w)$

## 4 Compute the Posterior

We want to compute the posterior  $p(w \mid X, Y)$ .

By Bayes rule, we have  $\frac{p(Y \mid X, w)p(w \mid x)}{p(Y \mid X)}$ .

We can rewrite

$$p(w \mid X, Y) = \frac{p(Y \mid X, w)p(w)}{\int_w p(Y \mid X, w)p(w)}.$$

$$\propto p(Y \mid X, w)p(w)$$

$$\propto p(Y \mid X, w)$$

Example: We imagine there is no conditioning variable. We just have a bunch of outputs,  $y = 0 \text{ or } 1$ , representing heads or tails if we consider coin flips. The parameter is  $\theta$  which is the probability of heads.

The first question is what is the likelihood.

$$p(y \mid \theta) = \theta^y(1 - \theta)^{1-y}$$

$$p(Y \mid \theta) = \prod_n \theta^{y_n}(1 - \theta)^{1-y_n} = \theta^{N_1}(1 - \theta)^{N_0}$$

We know the posterior is proportional to the likelihood and the prior.

Recall that

$$\theta_{\text{MLE}} = \frac{N_1}{N_1 + N_0}$$

Now, let's consider the prior  $p(\theta)$ . In theory, the prior could be anything. In practice, the Beta distribution is a convenient prior.

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}; \alpha, \beta > 0$$

If we choose  $p(\theta)$  is a Beta distribution, then

$$p(Y | \theta) \propto \theta^{N_1+\alpha-1}(1-\theta)^{N_0+\beta-1}$$

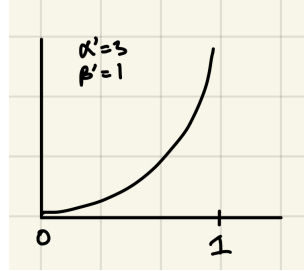
The Bernoulli and the Beta are conjugate. So by the form of the distribution, we have new beta distribution parameters:

$$\alpha' = N_1 + \alpha$$

and

$$\beta' = N_0 + \beta.$$

Let us apply what we found, starting with  $\alpha = \beta = 1$ . Say we have  $N_1 = 2, N_0 = 0$ . Then, our new beta distribution has  $\alpha' = 3$  and  $\beta' = 1$ . The distribution ends up looking like the following:



Previously we talked about MAP, which is the arg max solution over  $p(y | \theta)p(\theta)$  which is the arg max over  $\log p(Y | \theta) + \log p(\theta)$ .

As we add more data, the first term gets bigger but the second term always stays the same. Therefore, the second term doesn't have much effect.

For example, in this case,

$$\theta_{\text{MAP}} = \frac{N_1 + \alpha - 1}{N_1 + N_0 + \beta + \alpha - 2} = 1$$

and

$$\theta_{\text{MLE}} = 1.$$

## 5 Posterior Predictive

Now we want the probability that the next flip will come up heads. In the Bayesian perspective, we need to average over all our options.

$$\begin{aligned} p(y^* = 1 \mid y_1, \dots, y_n) &= \int_{\theta} p(y^* = 1 \mid \theta) p(\theta \mid y_1, \dots, y_n) \\ &= \int_{\theta} \theta p(\theta \mid y_1, \dots, y_n) \end{aligned}$$

This looks like the expectation of  $\theta$  under the posterior.

$$p(y^* = 1 \mid y_1, \dots, y_n) = \frac{\alpha + N_1}{\alpha + \beta + N_1 + N_0}$$

For our application, we get the posterior predictive  $p(y^* \mid Y) = \frac{3}{4}$ .

Compare this to  $p(y^* \mid \theta_{\text{MAP}}) = 1$  and  $p(y^* \mid \theta_{\text{MLE}}) = 1$ . When we averaged over all the possible  $\theta$ 's, we then get the posterior predictive of  $\frac{3}{4}$ .

If we had a different prior for  $\theta$ , we would have gotten a different posterior predictive.

## 6 Marginal Likelihood

$$\begin{aligned} p(Y = 1, \dots, y_n) &= \int_{\theta} \theta^{N_1} (1 - \theta)^{N_0} Z_{\alpha, \beta} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \\ &= \frac{Z_{\alpha, \beta}}{Z_{\alpha', \beta'}} \end{aligned}$$

Why is this useful?

### 6.1 Bayesian Occam's Razor aka Likelihood Regularization

In Bayesian inference, we start with a prior belief about the likelihood of different hypotheses being true. As we gather evidence, we update these beliefs using Bayes' theorem. Bayesian Occam's Razor comes into play when we compare different models or hypotheses to explain the observed data.

The principle suggests that when comparing models, we should consider not only how well they fit the data but also how complex they are. A more complex model may fit the data better, but it will also have more parameters or assumptions, making it more likely to overfit the data or be less generalizable to new data. Bayesian Occam's Razor provides a framework for balancing model complexity and goodness of fit. It favors simpler models that adequately explain the data over more complex ones, unless the additional complexity is justified by a significant improvement in explanatory power.

We can consider the Bayesian Occam's Razor as likelihood regularization because the goal is to prevent overfitting by favoring simpler models that generalize better to unseen data. The likelihood is providing the regularization, not the prior.

If we had a really complicated model class, then the probability will be spread quite thin, because probability will always add to 1. Most of our 50 dimensional choices are going to be crappy, a few choices will be good. But because there are so many models, the good ones won't do much.

## 7 Concept Check

Model 1:  $\hat{y} = a_0$

Model 2:  $\hat{y} = a_0 + a_1x$

Model 3:  $\hat{y} = a_0 + a_1x + a_2x^2$

Here is our prior:

$$p(a_0) = \text{Unif}[-100, 100]$$

$$p(a_1) = \text{Unif}[-1, 1]$$

$$p(a_2) = \text{Unif}[-100, 100]$$

Question: What is the posterior predictive at  $y = \frac{1}{2}$ ?

Solution is on the next page.

Model 1 cannot explain the data. Only one set of values that works for Model 2 and we end up with very little uncertainty. For model 3, we get a range of possible values.