WSABI Summer Research Lab **Summer 2025**

Lecture 11: Priors and the Power of Fake Data

Instructor: Jonathan Pipping *Author: Ryan Brill*

11.1 Motivating Example: End of Season Win Percentage

11.1.1 Problem Setup

Suppose an MLB team T has won W games and lost L games so far this season. How would you predict their end-of-season win percentage WP? There are some constraints on this problem:

- We have no access to team T's past or future schedule.
- We have no access to team T's previous season's data.

Together, these constraints rule out implementing regression models or controlling for strength of schedule. With this in mind, how can we guess their end of season win percentage? If you asked the average person, they'd probably guess their current win percentage:

$$\widehat{WP} = \frac{W}{W+L}$$

This is a reasonable guess, but there are some problems with it. For example, what if team T has only played a few games? If $W = 3$ and $L = 0$, then $\widehat{WP} = 1$. We clearly don't expect team T to win all 162 games they play, so we need to look for a better estimate.

11.1.2 Idea: Adding Fake Data

Suppose now that team T begins the season with W' fake wins and L' fake losses. Then a new guess for their end-of-season win percentage is:

$$\widehat{WP}' = \frac{W' + W}{(W' + W) + (L' + L)}$$

For concreteness, let's consider the example from earlier: $W = 3$ and $L = 0$. If we use Tom Tango's method of setting $W' = L' = 15$, then our adjusted estimate is:

$$\widehat{WP}' = \frac{15+3}{(15+3)+(15+0)} = \frac{18}{33} \approx 0.546$$

This is a very different estimate than $\widehat{WP} = 1$, but it's a much more stable one. What about later in the season? If team T has won $W = 45$ and lost $L = 30$ games, then our two estimates are:

$$\widehat{WP} = \frac{45}{45+30} = 0.6$$
$$\widehat{WP}' = \frac{15+45}{(15+45)+(15+30)} = \frac{60}{105} \approx 0.571$$

As the season progresses, the influence of the fake data is less and less, and $\widehat{WP}'$ and $\widehat{WP}$ get closer together. Still, which of these two estimates is better? We will have to formalize our estimates to answer this question.

11.1.3 Formalizing the Problem

Team T plays $n = 162$ games in a season. Suppose, for simplicity, that team T wins each game with probability p. Then $\{X_1, \ldots, X_n\}$ represent game outcomes, where

$$X_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \stackrel{d}{=} \text{Bernoulli}(p)$$

Suppose we have observed m games thus far in the season. Then our observed data is $\{X_1, \ldots, X_m\}$. Then our observed number of wins is

$$W = \sum_{i=1}^{m} X_i \sim \text{Binomial}(m, p)$$

since it is the sum of m i.i.d. Bernoulli(p) random variables. We then express the end-of-sesason win percentage as

$$WP \sim \frac{1}{n}\text{Binomial}(n, p)$$

Then we plan to use the observed data to estimate p (call it $\widehat{p}$), and then use $\widehat{p}$ to estimate WP as follows:

$$\widehat{WP} = \frac{1}{n}\mathbb{E}[WP] = \frac{1}{n}\mathbb{E}\left[\text{Binomial}(n, p)\right] = \frac{1}{n} \cdot n\widehat{p} = \widehat{p}$$

To accomplish this, we will explore the concept of the Maximum Likelihood Estimator (MLE).

11.2 Maximum Likelihood Estimator

11.2.1 Definition

The method of maximum likelihood estimation involves selecting the parameter estimate $\widehat{\theta}$ which maximizes the probability of observing the data we observed under the proposed model. To formalize this notion, we will first need to define the Likelihood function.

Definition 11.1 (Likelihood)**.** *If $\{X_i\}_{i=1}^n$ represents observed data drawn from a distribution $\mathbb{P}$ parametrized by parameter vector $\boldsymbol{\theta}$, then the Likelihood of observed data $\{X_i\}_{i=1}^n$ is defined as*

$$\mathcal{L}(\boldsymbol{\theta} \mid X_1, \ldots, X_n) = \mathbb{P}(X_1, \ldots, X_n \mid \boldsymbol{\theta})$$

In other words, the Likelihood represents the joint probability of observing the data we did conditional on the parameter(s) $\boldsymbol{\theta}$ of the underlying distribution. We are now ready to define the Maximum Likelihood Estimator (or MLE).

Definition 11.2 (Maximum Likelihood Estimator)**.** *If Θ represents the parameter space, $\{X_i\}_{i=1}^n$ represents the observed data, and $\mathcal{L}(\boldsymbol{\theta} \mid X_1, \ldots, X_n)$ represents the likelihood of data $\{X_i\}_{i=1}^n$, the Maximum Likelihood Estimator is the solution to the equation:*

$$\widehat{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta} \mid X_1, \ldots, X_n)$$

We'll now apply this method to our problem of predicting team T's end-of-season winning percentage.

11.2.2 Applying to Our Problem

Recall that $\{X_i\}_{i=1}^m$ are i.i.d. Bernoulli(p) random variables representing the observed games played by team T. Then by Definition 11.2, the Maximum Likelihood Estimator for p is given by:

$$\begin{aligned}\widehat{p}_{MLE} &= \arg\max_p \mathcal{L}(p \mid X_1, \ldots, X_m) \\ &= \arg\max_p \mathbb{P}(X_1, \ldots, X_m \mid p) \text{ from Definition 11.1} \\ &= \arg\max_p \prod_{i=1}^m \mathbb{P}(X_i \mid p) \text{ by independence of } \{X_i\}_{i=1}^m\end{aligned}$$

Recall that the probability mass function of a Bernoulli(p) random variable is given by $f(x) = p^x(1-p)^{(1-x)}$ for $x \in \{0, 1\}$. Then we have that

$$\begin{aligned}\widehat{p}_{MLE} &= \arg\max_p \prod_{i=1}^m p^{X_i}(1-p)^{1-X_i} \text{ since } X_i \sim \text{ Bernoulli}(p) \\ &= \arg\max_p p^{\sum_{i=1}^m X_i}(1-p)^{\sum_{i=1}^m (1-X_i)}\end{aligned}$$

Since we know that $\sum_{i=1}^m X_i = W$ and $\sum_{i=1}^m (1 - X_i) = L$, we have that

$$\begin{aligned}\widehat{p}_{MLE} &= \arg\max_p p^W (1-p)^L \\ &= \arg\max_p \log\left[p^W (1-p)^L\right] \text{ since log is monotonic} \\ &= \arg\max_p W\log(p) + L\log(1-p)\end{aligned}$$

To maximize this expression, we take the derivative with respect to p, set it equal to 0, and solve for p.

$$\begin{aligned}\frac{\partial}{\partial p}\left(W\log(p) + L\log(1-p)\right) = \frac{W}{p} - \frac{L}{1-p} &= 0 \\ \implies \frac{W}{p} &= \frac{L}{1-p} \\ \implies \widehat{p}_{MLE} &= \frac{W}{W+L}\end{aligned}$$

But wait, this is the same formula from earlier! The MLE is simply the observed win percentage through m games. Still,we know this is a bad estimate early in the season. Why did the MLE go wrong? How can we modify the MLE to get a better estimate?

Before, to improve our estimate of WP, we added some fake data (W', L') to the observed data (W, L), which allowed us to improve our estimate of WP to

$$\widehat{WP}' = \frac{W' + W}{(W' + W) + (L' + L)}$$

In adding this fake data, we used prior information: prior to the season, we assumed team T had W' wins and L' losses. We introduce a way to formalize this notion of prior information in the next section.

11.3 Introduction to Bayesian Statistics

11.3.1 Priors

To this point in the course, we have taken a **frequentist** approach to parameter estimation, which assumes a model where the data is random while the parameter is fixed. An alternative approach is the **Bayesian** approach, which assumes models in which we treat parameters as random variables with their own probability distributions.

At first, the difference between the two approaches may not be obvious, but the Bayesian approach introduces added flexibility in the model through the introduction of **priors**. What is a prior?

Definition 11.3 (Prior)**.** *A prior is a probability distribution over the parameter space Θ that represents our beliefs about the parameter before we observe any data.*

In our example, the addition of prior "fake" data essentially assigns a probability distribution to the parameter p which reflects our **prior** belief on what p is more likely to be! This allows for much more stable estimates of WP early in the season where the MLE is very unstable. Let's continue formalizing our win probability model using the **Beta-Binomial model**.

11.3.2 Beta-Binomial Model

Recall that in our original model with observed data $\{X_i\}_{i=1}^m$, we had that

$$W = \sum_{i=1}^{m} X_i \sim \text{Binomial}(m, p)$$

and that

$$\widehat{WP} = \frac{1}{n}\mathbb{E}[W] = \frac{1}{n}\text{Binomial}(n, p)$$

We will not change this portion of the model, but we **will** add a prior distribution to the parameter p, specifically a Beta distribution.

Definition 11.4 (Beta Distribution)**.** *The Beta distribution is characterized by two parameters, α and β, and has probability density function*

$$f(x \mid \alpha, \beta) = C \cdot x^{\alpha-1}(1-x)^{\beta-1} \text{ for } x \in [0,1], \alpha, \beta > 0$$

where C is a normalizating constant.

We visualize the Beta distribution for different values of α and β in Figure 11.1.

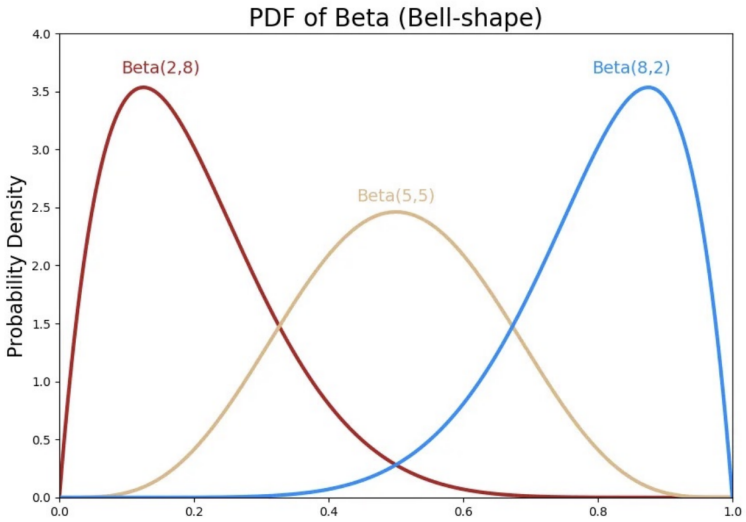

Figure 11.1: PDF of the Beta distribution for different values of α and β.

Note that the Beta distribution is very flexible in shape, but is always constrained between 0 and 1: this makes it a natural choice as a prior for our probability parameter p.

We define our Beta-Binomial model as follows:

$$\begin{cases} W \sim \text{Binomial}(m, p) \\ p \sim \text{Beta}(\alpha, \beta) \end{cases} \quad \text{where } \alpha = W' + 1 \text{ and } \beta = L' + 1$$

Now, as before, we wish to estimate p. A Bayesian approach to parameter estimation should incorporate both the prior information and the observed data, and the maximum a-posteriori (MAP) estimate is a natural way to do this.

11.3.3 Maximum a-Posteriori Estimate

Definition 11.5 (Maximum a-Posteriori (MAP) Estimate)**.** *If Θ represents the parameter space, $\{X_i\}_{i=1}^n$ represents the observed data, and $\mathbb{P}(\theta \mid X_1, \ldots, X_n)$ represents the posterior probability of parameter θ given the data, then the Maximum a-Posteriori (MAP) estimate is the solution to the equation:*

$$\widehat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} \mathbb{P}(\theta \mid X_1, \ldots, X_n)$$

In other words, the MAP estimate is the parameter value that maximizes the posterior probability of the parameter given the observed data.

Now let's apply this to our Beta-Binomial model to estimate p. Recalling that $W = \sum_{i=1}^m X_i$ encodes the observed data, we have that

$$\begin{aligned} \widehat{p}_{MAP} &= \arg\max_p \mathbb{P}(p \mid W) \text{ by Definition 11.5} \\ &= \arg\max_p \frac{\mathbb{P}(W \mid p)\mathbb{P}(p)}{\mathbb{P}(W)} \text{ by Bayes' Rule} \\ &= \arg\max_p \mathbb{P}(W \mid p)\mathbb{P}(p) \text{ since } \mathbb{P}(W) \text{ is constant w.r.t. } p \end{aligned} \tag{11.1}$$

Note that in this expression, $\mathbb{P}(W \mid p)$ is the likelihood of the data given the parameter p, and $\mathbb{P}(p)$ is the prior distribution of the parameter p. So then

$$\begin{aligned}\widehat{p}_{MAP} &= \arg\max_p \mathbb{P}(\text{Binomial}(m,p) = W)\mathbb{P}(\text{Beta}(\alpha,\beta) = p) \\ &= \arg\max_p \binom{m}{W} p^W (1-p)^{m-W} \cdot C \cdot p^{\alpha-1}(1-p)^{\beta-1}\end{aligned}$$

Removing constants with respect to p and noting that $m = W + L$, we have that

$$= \arg\max_p p^W (1-p)^L \cdot p^{\alpha-1}(1-p)^{\beta-1}$$

Then combining the exponents, we have that

$$= \arg\max_p p^{W+\alpha-1}(1-p)^{L+\beta-1}$$

From here, we follow the same procedure as before to maximize the expression: taking the log for simplicity, setting the derivative equal to 0, and solving for p. Following this procedure, we get that

$$\widehat{p}_{MAP} = \frac{(W+\alpha-1)}{(W+\alpha-1)+(L+\beta-1)}$$

If we let $\alpha = W' + 1$ and $\beta = L' + 1$, then we have that

$$= \frac{W' + W}{(W' + W) + (L' + L)}$$

So the MAP estimate is simply the win percentage if we add $\alpha - 1$ fake wins and $\beta - 1$ fake losses to the observed data! If we wanted to make informed choices of α and β, we might use data from previous seasons to tune these parameters.

11.3.4 Comparison of MLE and MAP

Note that if we were to set $\alpha = \beta = 1$, then the prior Beta distribution would be uniform over $[0, 1]$. In this case, the MAP estimate would be

$$\widehat{p}_{MAP} = \frac{W+1-1}{(W+1-1)+(L+1-1)} = \frac{W}{W+L} = \widehat{p}_{MLE}$$

So the MAP estimate is the same as the MLE when the prior is uniform. This is equivalent to setting up our model in the following way:

$$\begin{cases} W \sim \text{Binomial}(m,p) \\ p \sim \text{Beta}(1,1) \overset{d}{=} \text{Uniform}(0,1) \end{cases}$$

This is known as an **uninformative prior** which encodes no preference for any particular value of p. When we calculate the MAP estimate from Equation 11.1, we see that

$$\begin{aligned}
\widehat{p}_{MAP} &= \arg\max_p \mathbb{P}(p \mid W)\mathbb{P}(p) \\
&= \arg\max_p \mathbb{P}(p \mid W)\mathbb{P}(\text{Uniform}(0,1) = p) \\
&= \arg\max_p \mathbb{P}(p \mid W)(1) \\
&= \widehat{p}_{MLE} \\
&= \frac{W}{W+L}
\end{aligned}$$

11.3.5 Takeaways

- While frequentist statistics treats parameters (ex: p) as fixed, Bayesian statistics treats them as random variables with their own probability distributions.
- This allows us to blend observed data with prior information, outside information not seen in the data, to achieve more stable estimators and make better predictions.