The Power of Fake Data (Priors)

Q Suppose the Dodgers have won W and lost L games thus far in the season. How would you predict their end of season win percentage WP?

- 162 total games in the season
- no access to their schedule (e.g., ignore strength of schedule)
- Without using previous season's data (i.e. Regression).

Guess their end of season win percentage.

Naive guess (ask any rando on the street):

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

What's wrong with this?

When Dodgers have only played a few games, this estimate is bad.

Ex $W=3, L=0, \widehat{WP}=1$

Idea Add fake data.

Suppose the Dodgers begin the season with W' wins and L' losses.

New guess:

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

For concreteness:

$W=3, L=0, \quad \widehat{WP}=1$

$W=3, L=0, W'=15, L'=15, \quad \widehat{WP}' = \frac{18}{33} \approx .55$

quite different predictions early in the season

$W=45, L=30, \quad \widehat{WP} = \frac{45}{75} = .6$

$W=45, L=30, W'=15, L'=15, \quad \widehat{WP}' = \frac{60}{105} \approx .67$

similar predictions late in the season

Which is better?

Formalize this

Dodgers play $n = 162$ games in a season.

Suppose, for simplicity, that the Dodgers win each game with probability p.

Game outcomes $\{X_1, \ldots, X_n\}$, where

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

Suppose we have observed m games thus far in the season.

Observed data $\{X_1, \ldots, X_m\}$. Each X_i is 1 or 0.

Observed # wins $W = \sum_{i=1}^{m} X_i$.

So, $W \sim \text{Binomial}(m, p)$ $\quad m = \#\text{ trials (games)}$, $p = \text{prob. success (win)}$

and end-of-szn win percentage $WP \sim \frac{1}{n}\text{Binomial}(n, p)$

Idea use observed data to estimate p, call it $\hat{p}$

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

Maximum Likelihood estimate (MLE)

Choose $\hat{p}$ to be the value of p which maximizes the probability of observing the game outcomes $\{x_1, \ldots, x_m\}$ that we observed.

$$\hat{p}_{MLE} = \underset{p}{\operatorname{argmax}} \; \mathbb{P}(x_1, \ldots, x_m \mid p)$$

likelihood: $\mathbb{P}$(data given parameter)

$$= \underset{p}{\operatorname{argmax}} \; \mathbb{P}(x_1 \mid p) \cdot \mathbb{P}(x_2 \mid p) \cdot \ldots \cdot \mathbb{P}(x_m \mid p)$$

by independence

$$= \underset{p}{\operatorname{argmax}} \prod_{i=1}^{m} \mathbb{P}(x_i \mid p)$$

by def of product

$$= \underset{p}{\operatorname{argmax}} \prod_{i=1}^{m} p^{x_i} (1-p)^{1-x_i}$$

because $x_i \sim \text{Ber}(p)$

$x_i = 1$ means $p^{x_i}(1-p)^{1-x_i} = p$

$x_i = 0$ means $p^{x_i}(1-p)^{1-x_i} = 1-p$

$$= \operatorname*{argmax}_p \; p^{\sum_{i=1}^{m} x_i} (1-p)^{\sum_{i=1}^{m} (1-x_i)}$$

$$= \operatorname*{argmax}_p \; p^W (1-p)^L$$

where $W = \sum_{i=1}^{m} x_i$ = number of wins (ones)

$L = \sum_{i=1}^{m} (1-x_i)$ = number of losses (zeros)

$$= \operatorname*{argmax}_p \; \log \left[p^W \cdot (1-p)^L \right]$$

because log is monotonic increasing
to maximize $f(p)$ is to maximize $\log f(p)$

$$= \operatorname*{argmax}_p \; W \log p + L \log (1-p)$$

to maximize the function $p \mapsto W \log p + L \log(1-p)$
take the derivative and set it equal to 0
(and check that the 2nd derivative is negative).

$$\frac{d}{dp} \left[W \log p + L \log (1-p) \right]$$

$$= W \cdot \frac{1}{p} - L \cdot \frac{1}{1-p} = 0$$

$$\Rightarrow \frac{W}{p} = \frac{L}{1-p} \Rightarrow p = \frac{W}{L}(1-p)$$

$$\Rightarrow p\left(1+\frac{W}{L}\right) = \frac{W}{L} \Rightarrow p = \frac{\frac{W}{L}}{1+\frac{W}{L}}$$

$$\Rightarrow \hat{p}_{MLE} = \frac{W}{W+L}$$

same formula from earlier!!

The MLE is simply the observed win percentage midway through the season!

But we know this is a bad estimate early in the season.

So, why did the MLE go wrong?? How do we add the fake data W', L' to the MLE to get $\frac{W+W'}{W+W'+L+L'}$??

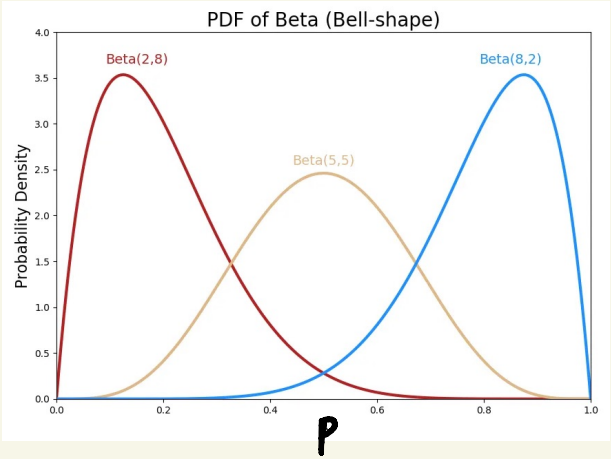
Before, to improve our estimate of WP, we added some fake number of wins W' and some fake number of losses L' to the Dodgers' record.

In other words, we assumed **Prior Information**: prior to the season, assume the Dodgers have W' wins and L' losses.

Formally, we use the **Beta-Binomial model**:

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

Beta distribution has density $f(p|\alpha,\beta) = C \cdot p^{\alpha-1}(1-p)^{\beta-1}$ on the interval $p \in [0,1]$, where C is a constant chosen so that the distribution integrates to 1.



For example, $p \sim \text{Beta}(5,5)$ encodes a preference that p is closer to 0.5

As before, we wish to estimate p, this time with a Maximum a-Posteriori (MAP) Estimate:

$$\hat{p}_{MAP} = \underset{p}{\text{argmax}} \; \underbrace{P(p|w)}_{\text{Posterior} = P(\text{parameter}|\text{data})}$$

$$= \underset{p}{\text{argmax}} \; \frac{P(w|p) \cdot P(p)}{P(w)} \quad \text{by Bayes' Rule}$$

$$= \underset{p}{\text{argmax}} \; \underbrace{P(w|p)}_{\text{likelihood}} \cdot \underbrace{P(p)}_{\text{prior}}$$

Since $P(w)$ has no p term

$$= \underset{p}{\text{argmax}} \; \mathbb{P}\left(\text{Binomial}(m,p) = w\right) \cdot \mathbb{P}\left(\text{Beta}(\alpha,\beta) = p\right)$$

$$= \underset{p}{\text{argmax}} \; \binom{m}{w} p^{w} (1-p)^{m-w} \cdot C \, p^{\alpha-1} (1-p)^{\beta-1}$$

$$= \underset{p}{\text{argmax}} \; p^{w}(1-p)^{L} \cdot p^{\alpha-1}(1-p)^{\beta-1}$$

Since the constants $\binom{m}{w}$ and C don't have p terms

$$= \underset{p}{\text{argmax}} \; p^{w+\alpha-1} (1-p)^{L+\beta-1}$$

$$= \cdots \text{ same process as before}$$

$$= \frac{w+\alpha-1}{w+\alpha-1+L+\beta-1}$$

$$= \frac{w+w'}{w+w'+L+L'} \quad \text{if} \quad \begin{aligned} w' &= \alpha-1 \\ L' &= \beta-1 \end{aligned}$$

The MAP estimate is simply the win percentage if we add $\alpha-1$ fake wins and $\beta-1$ fake losses!!

Note: $\alpha=1, \beta=1 \Rightarrow \hat{p}_{MAP} = \hat{p}_{MLE}$

add no fake data

Can use past seasons to tune a smart choice for α, β.

HW: use past seasons' data to find good values of α and β, assuming $\alpha=\beta$ (i.e. assuming each time has a prior with percentage of 50%)

Tom Tango: $W'=L'=15$ is good

HW: use past seasons' data and pre-season win totals to tune smart values of α and β for each team

Takeaway: blending observed data with prior knowledge is often a smart way to make better predictions which don't overfit to random noise