A high-level overview of Bayesian Statistics

Bayesian Idea: Treat a parameter as having an unknown Distribution to be estimated, rather than as an unknown fixed number to be estimated.

Ex 1 Predict end-of-season win percentage from mid-season Wins and Losses.

Beta-Binomial model

$$\begin{cases} W \sim \text{Binomial}(n, p) \\ p \sim \text{Beta}(\alpha, \beta) \end{cases}$$

When we model the latent team's win probability p using a Beta Prior, we encode the prior information that p is more likely to be near 1/2 (say, in .3,.7) than to be very near 0 or 1.

Then, we found using Bayes Rule that the posterior distribution $p|W,L$ is

$$p|W,L \sim \text{Binomial}\left(n+\alpha+\beta-2, \frac{W+\alpha-1}{W+L+\alpha+\beta-2}\right)$$

and the Bayes estimate is the posterior mean

$$\hat{p}^{(\text{Bayes})} = \mathbb{E}[p|W,L] = \frac{W+(\alpha-1)}{W+L+(\alpha+\beta-2)}.$$

Ex 2 Predict end-of-season batting average from mid-season batting average and number of at-bats.

Normal-Normal model

i = player i

H_i = # hits, N_i = # at-bats, $X_i = \frac{H_i}{N_i}$

$X_i \sim N(\mu_i, \sigma_i^2)$

$\sigma_i^2 = C/N_i$, C known

$\mu_i \sim N(\mu, \tau^2)$

When we model player i's latent quality μ_i using a Normal Prior, we encode the prior information that player i is also a baseball player, allowing us to share strength across players.

Then, we found using Bayes Rule that the posterior distribution $\mu_i | X$ is

$$\mu_i | X \sim N\left(\frac{\left(\frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)}, \frac{1}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)} \right)$$

and the Bayes estimate is the posterior mean

$$\hat{\mu}_i^{(\text{Bayes})} = \frac{\frac{X_i}{\sigma_i^2} + \frac{\hat{\mu}}{\hat{\tau}^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\hat{\tau}^2}}$$

using $\hat{\mu}$ and $\hat{\tau}^2$ since μ, τ are unknown.

Ex 3 **Bayesian Regression**

Suppose we have a regression model $y = X\beta + \varepsilon$ with mean zero noise $\mathbb{E}\varepsilon = 0$.

For instance, our park effects model (so we want to estimate the park effects β), or a power score model like Bradley Terry.

A Bayesian puts a prior distribution on the parameters β (and, a distribution on the error term ε) to end up with a posterior distribution $\beta | X$ or a posterior mean $\mathbb{E}[\beta | X]$, quantifying our best guess of β after seeing the data X, while also incorporating our prior information.

* "Standard" Regression setup:

$$y = X\beta + \varepsilon$$

$$\left.\begin{array}{l} \text{Homoscedasticity} \longrightarrow \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\ \qquad\qquad\qquad\quad \Rightarrow \varepsilon \sim N(0, \sigma^2 I) \end{array}\right\} \Rightarrow y \sim N(X\beta, \sigma^2 I)$$

* Bayesian Regression ex 1: uninformative Uniform Prior $\quad \beta \propto 1$

* Bayesian Regression ex 2: Normal Prior $\quad \beta \sim N(0, \lambda)$

Bayesian Regression v1

$$\begin{cases} y \sim N(X\beta, \sigma^2 I) \\ \beta \propto 1 \quad \text{uniform PRIOR} \\ \text{Find posterior dist } \beta | X \end{cases}$$

$$\mathbb{P}(\beta|y) = \frac{\mathbb{P}(y|\beta)\,\mathbb{P}(\beta)}{\mathbb{P}(y)} \quad \text{by Bayes Rule}$$

$$\propto \mathbb{P}(y|\beta)\cdot\mathbb{P}(\beta) \quad \text{since } \mathbb{P}(y) \text{ doesn't depend on } \beta$$

$$\propto \mathbb{P}(y|\beta) \quad \text{since } \mathbb{P}(\beta) \propto 1$$

$$\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right).$$

Bayes estimate of β is to choose the β which maximizes the posterior probability

Maximum A-Posteriori (MAP):

$$\hat{\beta}^{(Bayes)} = \underset{\beta}{\operatorname{argmax}}\ \mathbb{P}(\beta|y)$$

$$= \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 = \hat{\beta}^{(OLS)} = (X^TX)^{-1}X^Ty.$$

* With an uninformative uniform prior, Bayesian Regression is the same as Ordinary least Squares!

Bayesian Regression v2

$$\begin{cases} y \sim N(X\beta, \sigma^2 I) \\ \beta \sim N(0, \frac{1}{\lambda} I) \\ \text{Find posterior dist } \beta | X \end{cases}$$

$$\mathbb{P}(\beta|y) = \frac{\mathbb{P}(y|\beta)\,\mathbb{P}(\beta)}{\mathbb{P}(y)}$$ by Bayes Rule

$$\propto \mathbb{P}(y|\beta)\cdot\mathbb{P}(\beta)$$ since $\mathbb{P}(y)$ doesn't depend on β

$$= \mathbb{P}(N(X\beta, \sigma^2 I) = y)\cdot \mathbb{P}(\beta = N(0,\lambda))$$

$$\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\frac{(x_i^T\beta - y_i)^2}{\sigma^2}\right) \cdot \prod_{j=1}^{p} \exp\left(-\frac{1}{2}\frac{(\beta_j - 0)^2}{1/\lambda}\right)$$

$$= \exp\left(-\frac{1}{2}\left[\lambda\sum_{j=1}^{p}\beta_j^2 + \sum_{i=1}^{n}\frac{(x_i^T\beta - y_i)^2}{\sigma^2}\right]\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\left[\lambda\sigma^2\sum_{j=1}^{p}\beta_j^2 + \sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right]\right).$$

Bayes estimate of β is to choose the β which maximizes the posterior probability Maximum A-Posteriori (MAP):

$$\hat{\beta}^{(Bayes)} = \underset{\beta}{\operatorname{argmax}}\ \mathbb{P}(\beta|y) = \underset{\beta}{\operatorname{argmax}}\ \lambda\sum_{j=1}^{p}\beta_j^2 + \sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

(letting $\lambda \leftarrow \frac{\lambda}{\sigma^2}$) $$= \hat{\beta}^{(Ridge)} = (X^TX + \lambda I)^{-1}X^Ty.$$

* With a NORMAL PRIOR, Bayesian Regression is Ridge Regression. This makes perfect sense: $\beta \sim N(0, \lambda I)$ encourages β to be closer to 0. λ controls by how much, and that is the same as penalizing β!

* So far we have mainly focused on finding the posterior mean, i.e. our best estimate of the parameters after seeing the data.

* To make decisions in sports (e.g. player valuation or play selection), we need not only know our best estimate of the value of the player/play/decision, but also uncertainty quantification (e.g., error bars or prediction intervals) to describe how confident/certain we are about a value.

Ex Suppose we have a prediction of end-of-season win percentage to be 74%.
A 95% prediction interval of [72%, 76%] is much different from [60%, 88%] and from [0%, 100%].

Bayesian Idea Estimate the FULL Posterior Distribution of an unknown parameter. This gets us error bars (uncertainty quantification), with which we can create more complete decisions.

* In a general Bayesian statistical model, the parameters don't all have Gaussian priors or likelihoods, and we can't write on paper a closed-form analytical solution for the posterior distribution.

* Typically need to <u>approximate</u> the posterior distribution using MCMC sampling methods like

  {Gibbs sampling
  Hamiltonian Monte Carlo
  No U-Turn Sampling,

  which we won't cover here (take Shane's class), using a probabilistic programming language

  [Stan
  Jags
  NumPyro

* Next Time: a full example of Bayesian modeling in sports.

Takeaway: Create a fully Bayesian probability model, which allows quantify both the effect size and uncertainty of a phenomenon of interest by approximating the full posterior distribution.