Lecture 7

The Beta Distribution

Consider the dataset $x =< 0, 0, 0 >$. Then, $\hat{\theta_{MLE}} = 0$. Furthermore,

$$\theta \sim U(0,1) \rightarrow \theta|x = Beta(\Sigma x_i + 1, n - \Sigma x_i + 1)$$

$$\hat{\theta}_{MMSE} = E[\theta|x] = \frac{\Sigma x_i + 1}{n+2} = \frac{1}{5}$$

$$\hat{\theta}_{MMAE} = Med[\theta|x] = qbeta(0.5, 1, 4) = 0.1591$$

$$\hat{\theta}_{MAP} = \frac{\Sigma x_i + 1 - 1}{n+2-2} = \frac{0}{3} = 0$$

This all makes sense based on what we have learned. Remember, the probabilities of all $\theta \in \Theta$ are uniformly distributed, and defined by the Beta function.

$$P(\theta) = U(0,1) = Beta(1,1)$$

$$P(\theta|X_1) = \frac{P(X_1|\theta}{P(X_1)} = Bern(1,2)$$

$$P(\theta|X_2) = \frac{P(X_2|\theta}{P(X_2)} = Bern(1,3)$$

$$P(\theta|X_3) = \frac{P(X_3|\theta}{P(X_3)} = Bern(1,4)$$

Where the probability of theta is updated with new data with every iteration. The beta prior yields a beta posterior for the parameter model $\mathcal{F} : iidBern(\theta)$. Lets prove this generally:

$$\mathcal{F} : iidBern(\theta),\ P(\theta) = Beta(\alpha, \beta)$$

We'll simply churn through the Bayesian formula.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$P(X) = \int_0^1 P(X|\theta)P(\theta)\ d\theta$$

Simplifying,

$$P(\theta|X) = \frac{\theta^{\Sigma X_i}(1-\theta)^{n-\Sigma X_i}\ \frac{1}{(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 P(X|\theta)P(\theta)\ \frac{1}{(\alpha,\beta}\theta^{\alpha-1}(1-\theta)^{\beta-1}}$$

This simplifies further to,

$$Beta(\alpha,\beta) = Beta(\alpha+\Sigma X_i, \beta+n-\Sigma X_i)$$

$$Beta(\alpha,\beta) = P(\theta)$$

$$Beta(\alpha+\Sigma X_i, \beta+n-\Sigma X_i) = P(\theta|X)$$

Definition: Conjugacy
The prior and the posterior are the same random variable. We say that the beta, β, it the *conjugate prior* for the iid Bernoulli likelihood model. α and β are parameters of the prior distribution. Because they're a step removed from parameters, θ, the target of our inference, they are called hyper parameters. Who specifies their values? The statistician does! If you claim an alpha and beta, you have made your prior explicit.

We are now going to probe that $\mathcal{F}$: $iidBern(\theta)$ is the same as $\mathcal{F}$: one realization of a $Binomial(n,\theta)$ with n fixed.

$$X_1,\ ...,\ X_n \sim iid\ Bern(\theta) \rightarrow \Sigma X_i \sim Binom(n,\theta)$$

We let X represent the sum of successes.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int_0^1 P(X|\theta)P(\theta)d\theta}$$

$$= Beta(\alpha+\Sigma X_i, \beta+n-\Sigma X_i) = P(\theta|X)$$

The beta is the conjugate prior for the binomial likelihood model. $\alpha+\Sigma X_i$, represents the prior, and $\beta+n-\Sigma X_i)$ represents the posterior.

$$Beta(\alpha,\beta) = Beta(\alpha+\Sigma X_i, \beta+n-\Sigma X_i)$$

Given this, we can find descriptive estimates.

$$\hat{\theta}_{MMSE} = \frac{x+\alpha}{n+\alpha+\beta}$$

$$\hat{\theta}_{MAP} = \frac{x+\alpha-1}{n+\alpha+\beta-2}$$

$$\hat{\theta}_{MMAE} = \hat{\theta}(0.5, x+\alpha, n-x+\beta)$$

$$Beta(\alpha,\beta) = Beta(\alpha+\Sigma X_i, \beta+n-\Sigma X_i)$$

$x = successes$
$n-x = failures$

$\alpha = pseudo - successes$
$\beta = pseudo - failures$
$\alpha + \beta = n_0 = pseudo - trials$

Laplace's principle of indifference with prior is $\theta \sim U(0,1) = Beta(1,1)$ which means $\alpha = 1$ and $\beta = 1$ which means you are pretending to see 2 pseudo-trials where 1 is a pseudo-success and 1 is a pseudo-failure. $E[\theta] = 0.5$.

Consider our MMSE Bayesian point estimate, $\hat{\theta}_{MMSE} = \frac{x+\alpha}{n+\alpha+\beta}$.

$$= \frac{x}{n+\alpha+\beta} \times \frac{n}{n} + \frac{\alpha}{n+\alpha+\beta} \times \frac{\alpha+\beta}{\alpha+\beta}$$

$$= \frac{n}{n+\alpha+\beta} \times \frac{x}{n} + \frac{\alpha+\beta}{n+\alpha+\beta} \times \frac{\alpha}{\alpha+\beta}$$

Note:
$\frac{x}{n} = \hat{\theta}_{MLE}$
$\frac{\alpha+\beta}{n+\alpha+\beta} = \rho$
$\frac{\alpha}{\alpha+\beta} = E[\theta]$
$\frac{n}{n+\alpha+\beta} = 1 - \rho$

$$= (1-\rho)\hat{\theta}_{MLE} + (\rho)E[\theta]$$

Above we have defined a linear combination of the MLE and prior mean. This means that the MMSE in the beta binomial conjugate model is a shrinkage estimator. It takes the MLE and it shrinks it towards the prior mean. Notice that the prior is effected at a rate of $\frac{1}{n}$, the more data points there are, the less the prior matters.

$$lim_{n\to\infty}\ \rho = 0$$

Why do we do this though? What is the purpose of "shrinking" the MLE towards the prior mean? We will discuss this at a later time. Thus far, we have only talked about the first goal of inference, i.e., point estimation. What about the second goal, that of confidence sets. How do we provide a region of reasonable values of θ?

Consider the following, $x = 1, n = 2, \alpha = \beta = 1 \rightarrow = (2,2)$.

Lets say I want a set R such that $P(\theta in R|x) = 1 - \alpha_0$. where R represents the 'middle' of the posterior distribution, where our confidence sit centers itself on the distribution and consider 95% of all intervals. By stripping off $\frac{\alpha_0}{2}$ from the tail ends of the distribution, the remaining central portion of the distribution is where our *credible region* (CR) for θ at level $1 - \alpha_0$:

$$CR_{\theta,1-\alpha_0} = [Quantile[\alpha_0|x, \frac{\alpha_0}{2},\ Quantile[\alpha_0|x, 1 - \frac{\alpha_0}{2}]]$$

This further simplifies into the beta-binomial model, defined as so,

$$= [qbeta(\frac{\alpha_0}{2}, \alpha + x, \beta + n - x), qbeta(1 - \frac{\alpha_0}{2}, \alpha + x, \beta + n - x)]$$