Stat210A: Theoretical Statistics **Lecture Date: Sept. 30, 2014**

Bayesian Inference

Lecturer: Tamara Broderick *Scribe: Ross Boczar*

1 Being Bayesian

Bayes' Theorem (for densities): X, Θ are random variables

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

- Joint $p(x, \theta)$
- Marginal $p(\theta)$
- Conditionals $p(x|\theta), p(\theta|x)$
- Want to use these to do inference

One school of thought: statistics $\iff$ "inverse probability"

- Probabilistic model/simulation: fixed $\theta \to$ random X
- Statistical inference: observe $x \to$ knowledge of Θ

Bayesian thinking:

- Represent knowledge of Θ with a distribution: $p(\theta)$ (prior); $p(\theta|x)$ (posterior)

Example

Θ: true proportion of U.S. residents that watch *Game of Thrones*

X: # who watch out of n residents

Bayesian model:

$$p(x|\theta) = \text{Bin}(x|n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$
$$p(\theta) = \text{Beta}(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

For the prior distribution, a, b are known as *hyperparameters*. Note that when $a = b = 1$, we have a uniform distribution. In general, a, b changing can represent us having more information about θ. Some useful identities:

$$\begin{aligned} E[\Theta] &= \frac{a}{a+b} \\ \text{Var}(\Theta) &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

We then have (using our knowledge of the Beta distribution to find the normalizing constant):

$$\begin{aligned} p(\theta|x) &\propto_\theta p(x|\theta)p(\theta) \\ &\propto_\theta \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ \implies p(\theta|x) &= \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}\theta^{x+a-1}(1-\theta)^{n-x+b-1} \\ &= \text{Beta}(\theta|x+a, n-x+b). \end{aligned}$$

2 Frequentist Analysis

Bayes: calculate posterior

Frequentist: Analyze estimator

Can get estimator from posterior (e.g. MAP)

Theorem 1. *(Keener Theorem 7.1) Assume $L(\theta, d(x)) \geq 0$, $\forall\, \theta \in \Omega, \forall d$. Then if there exists δ_0 such that $EL(\Theta, \delta_0(X)) < \infty$, and for a.e. x there exists a value $\delta_\Lambda(x)$ minimizing $E[L(\Theta, d(X))|X = x]$ with respect to d, then δ_Λ is a* Bayes estimator.

Quadratic loss

Assume quadratic loss, e.g. $L(\theta, d) = (d(x) - g(\theta))^2$

$$\begin{aligned} \rho &\equiv E[L(\Theta, \delta_\Lambda(X))|X = x] \\ &= \int (d(x) - g(\theta))^2 p(\theta|x)d\theta \\ &= d^2(x) - 2d(x)\int g(\theta)p(\theta|x)d\theta + \int g^2(\theta)p(\theta|x)d\theta \end{aligned}$$

Setting the derivative with respect to $d(x)$ to zero gives

$$\begin{aligned} 0 &= 2d(x) - 2\int g(\theta)p(\theta|x)d\theta \\ \implies \delta_\Lambda(x) &= \int g(\theta)p(\theta|x)d\theta = E[g(\Theta)|X = x]. \end{aligned}$$

When $g(\theta) = \theta$, this value is the posterior mean.

Example 2. $L(\theta, d) = |d(x) - \theta| \implies \delta_\Lambda(x) =$ posterior median.

Game of Thrones continued

Assume quadratic loss with $g(\theta) = \theta$. Then:

$$\begin{aligned}\Theta|X = x &\sim \text{Beta}(x + a, n - x + b)\\ \delta_\Lambda(x) = E[\theta|X = x] &= \frac{x + a}{n + a + b}\\ &= \left(\frac{n}{n + a + b}\right)\left(\frac{x}{n}\right) + \left(\frac{a + b}{n + a + b}\right)\left(\frac{a}{a + b}\right).\end{aligned}$$

We see that the posterior mean is a weighted average of the prior mean and the MLE of the likelihood.

Remark 3. There is a posterior mean decomposition.

Remark 4. a, b behave like "extra" or "prior" data.

Remark 5. Calculating the posterior was easy.

We can iterate on this concept of updating beliefs:

$$\begin{aligned}\Theta \sim \text{Beta}(a, b),\ X_1|\theta &\sim \text{Bin}(n_1, \Theta)\\ \Theta|X_1 \sim \text{Beta}(x_1 + a, n_1 - x_1 + b),\ X_2|\Theta &\sim \text{Bin}(n_2, \Theta)\\ \implies \Theta|X_1, X_2 &\sim \text{Beta}(x_1 + x_2 + a, (n_1 - x_1) + (n_2 - x_2) + b).\end{aligned}$$

We can also get this directly from the factorization of the joint density, assuming the X_i are i.i.d. conditional on Θ

3 Conjugacy

Definition 6. A family of distributions is *conjugate* to a likelihood if, for any prior in the family, the posterior is also in the family.

- The set of all probability distributions is conjugate to any likelihood...
- Beta priors are conjugate to the binomial likelihood.

Ex.

$$p(x|\theta) \sim \text{Poisson}(x|\theta) \propto_\theta \theta^x e^{-x}; \quad p(x_{1:n}|\theta) \propto_\theta \theta^{\sum x_i} e^{-n\theta} \quad \text{(assuming i.i.d.)}$$

A good guess for a conjugate prior would then be $p(\theta) \propto_\theta \theta^{a-1} e^{-b\theta}$, which is a Gamma distribution with parameters a, b. Thus, we have that

$$p(\theta|x_{1:n}) \propto_\theta \theta^{a + \sum x_i - 1} e^{-(b+n)\theta}$$

So

$$p(\theta|x_{1:n}) = \text{Gamma}(\theta|a + \sum x_i, b + n).$$

Therefore, under squared loss, the Bayes estimator (the posterior mean) is simply

$$\begin{aligned}\delta_\Lambda(x) &= \frac{a + \sum x_i}{b + n}\\ &= \left(\frac{b}{b + n}\right)\left(\frac{a}{b}\right) + \left(\frac{n}{b + n}\right)\left(\frac{\sum x_i}{n}\right),\end{aligned}$$

which is again a weighted average of the prior mean and the MLE.

Conjugacy for exponential families

$$
\begin{aligned}
p(x|\eta) &= h(x)\exp(\eta^T T(X) - A(\eta)) \\
p(x_1, \ldots, x_n|\eta) &= \left(\prod h(x_i)\right)\exp(\eta^T \sum T(x_i) - nA(\eta))
\end{aligned}
$$

Conjugate prior:

$$p(\eta) = \exp(\tau^T \eta - n_0 A(\eta) - \tilde{A}(\tau; n_0)).$$

Posterior:

$$p(\eta|x_1, \ldots, x_n) \propto_\eta \exp((\tau + \sum T(x_i))^T \eta - (n + n_0)A(\eta))$$

What about the posterior mean?

Let $E\mu = EE[T(x)|\eta] = E\nabla_\eta A(\eta)$. Also note that $\nabla_\eta p(\eta) = p(\eta)(\tau - n_0 \nabla_\eta A(\eta))$.

Then, we have that:

$$
\begin{aligned}
\int p(\eta)(\tau - n_0\nabla_\eta A(\eta))d\eta &= \int \nabla_\eta p(\eta) d\eta = 0 \quad \text{(by Green's theorem)} \\
&= \tau - n_0 E\nabla_\eta A(\eta) \\
\implies E\nabla_\eta A(\eta) &= \frac{\tau}{n_0} = E\mu.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
E[\mu|X_1 = x_1, \ldots, X_n = x_n] &= \frac{\tau + \sum T(x_i)}{n + n_0} \\
&= \left(\frac{n}{n+n_0}\right)\left(\frac{\sum T(x_i)}{n}\right) + \left(\frac{n_0}{n+n_0}\right)\left(\frac{\tau}{n_0}\right).
\end{aligned}
$$

Again we have the posterior mean decomposition.