STATISTICS
TEXAS A&M UNIVERSITY

Basics of Bayesian Statistics

October 28, 2015

Principle of Frequentist Inference

Probabilities describe long run relative frequency. Only calculate probabilities of repeatable events.

- random variables $X_1, \ldots, X_n$ are iid with density $p(x|\theta)$
- the unknown parameter θ is some fixed value
  - example: $X_1, \ldots, X_{100} \sim N(\theta, 1)$
- probabilistic statements refer to X_i at some value of θ

$$P_{\theta=0}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > 0.1\right) \approx 0.16$$

- do not compute probabilities of θ being somewhere

$$P(\theta > 1) = \ ???$$

Bayes Theorem (Not Bayesian Statistics)

Bayes Theorem is a result from probability theory used by both Bayesian and frequentist statisticians.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Bayesian Idea

Uncertainty about non–repeatable events (eg the value of parameters) can be described by probabilities.

- random variables $X_1, \dots, X_n$ are iid with density $p(x|\theta)$
- the possible values for θ are summarized by a prior $\pi(\theta)$ $\pi \neq 3.14$ here
  - $\pi(\theta) > 0 \, \forall\theta$, $\int \pi(\theta) = 1$
  - π represents prior (before seeing the data) belief about θ

The posterior (belief about parameter after seeing data):

$$p(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)} = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta}$$

Note: $p(X|\theta)$ and $\pi(\theta)$ are known, but $p(\theta|X)$ may be hard to compute.

Conjugate Family

- Under special conditions, the likelihood ($p(\theta|\vec{X})$) and the prior ($\pi(\theta)$) are conjugate, meaning the posterior has the same form as the likelihood.
- In such cases, computing the posterior is easy.
- While priors ideally should be chosen to represent prior belief, often they are chosen to be conjugate.

Normal Example

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2) = p(x|\mu)$ (assume σ^2 is known)
- $\mu \sim N(\mu_0, \sigma_0^2) = \pi(\mu)$

The posterior is

$$
\begin{aligned}
p(\mu|\vec{X}) &= \frac{p(\vec{X}|\mu)\pi(\mu)}{p(\vec{X})} \\
&\propto p(\vec{X}|\mu)\pi(\mu) \\
&\propto exp(-\sum(x_i - \mu)^2/(2\sigma^2))exp(-(\mu - \mu_0)^2/(2\sigma_0^2)) \\
&\propto exp\left(\frac{-(\mu^2(n\sigma_0^2 + \sigma^2) - 2\mu(\sigma_0^2 \sum X_i + \mu_0\sigma^2))}{2\sigma_0^2\sigma^2}\right) \\
&\propto exp\left(\frac{-\left(\mu - \frac{\sigma_0^2 \sum X_i + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}\right)^2}{2\left(\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}\right)}\right)
\end{aligned}
$$

Normal Example

So

$$p(\mu|\vec{X}) = N\left(\frac{\sigma_0^2 \sum X_i + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}, \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}\right)$$

- The posterior has the same form as the likelihood (both normal), so this is a conjugate family.
- The posterior represents your beliefs about the parameter after having seen the data.

Bayesian Point Estimators

Once you have a posterior, you may want to summarize it with a point estimate of θ.

Common Point Estimators:

- maximum–a–posteriori estimator: $\widehat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|X)$
- posterior mean: $\widehat{\theta}_{M} = \int \theta p(\theta|X) d\theta$

Bayesian Point Estimators (Normal Example)

The mean of a normal equals the mode of the normal, so

$$\widehat{\theta}_{MAP} = \widehat{\theta}_M = \frac{\sigma_0^2 \sum X_i + \mu_0 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

- when n is large

  $$\approx \frac{1}{n} \sum X_i$$

  the prior "washes out."
- if n is 1

  $$= \frac{\sigma_0^2 X + \mu_0 \sigma^2}{\sigma_0^2 + \sigma^2}$$

  weighted average between the prior and the data

Bayesian (Credible) Intervals / Regions

An $100\alpha\%$ credible interval is any interval $[L, U]$ such that

$$\alpha = \int_L^U p(\theta|\vec{X})d\theta$$

- This is the Bayesian version of the confidence interval.
- For normals, about 95% of the data is within 2 standard deviations of the mean. So a 95% credible interval for the normal example is

$$\frac{\sigma_0^2 \sum X_i + \mu_0 \sigma^2}{n\sigma_0^2 + \sigma^2} \pm 2\sqrt{\frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}}$$

Bayesian Computation Preview

- fully conjugate models are more the exception than the rule
- often there is no closed form solution for $p(\theta|\vec{X})$
- techniques such as Markov Chain Monte Carlo (MCMC) are used to draw samples

$$\theta_1, \ldots, \theta_m \sim p(\theta|\vec{X})$$

- point estimators and credible intervals are constructed from $(\theta_1, \ldots, \theta_m)$

Schedule for Next 2 Weeks

- October 29:
  - MCMC for Bayesian Intrinsic Scatter Regression Model
  - Discuss Problem 2 of Project 3
- November 3:
  - Neural Networks in Source Extractor
  - Model Checking
- November 5:
  - Techniques in Supernovae Search
  - Model Checking
- November 10:
  - Project 3 Due
  - Start Extragalatic Astronomy