Lecture 7:

Bayesian Inference

The Bayesian Philosophy

The Bayesian Philosophy

The statistical methods we’ve discussed so far are known as **frequentist** (or **classical**) methods. Frequentist point of view is:

1. Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.

2. Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statement can be made about parameters.

3. Statistical procedures should be designed to have well-defined long run frequency probabilities. For example, a 95% confidence interval should trap the true value of the parameter with limiting frequency at least 95%.

The Bayesian Philosophy

But there is another approach to inference called **Bayesian inference**. Bayesian point of view is:

1. Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data subject to random variation – for example, “the probability that Einstein drank a cup of tea on August 1, 1948 is 0.35”. This does not refer to any limiting frequency.

2. We can make probability statements about parameters, even though they are fixed constants.

3. We make inferences about a parameter θ by producing a probability distribution for it. Inferences, such as point/interval estimates are extracted from this distribution.

The Bayesian Philosophy

- Bayesian inference is a controversial approach – it inherently embraces a subjective notion of probability. Bayesian methods provide no guarantees on long run performance.

- Statistics puts more emphasis on frequentist methods although Bayesian methods certainly have a presence – certain machine learning communities embrace Bayesian methods very strongly.

- We’ll discuss the strengths and weaknesses of the Bayesian approach.

The Bayesian Method

The Bayesian Method

Bayesian inference is usually carried out as follows:

1. We choose a probability distribution $f(\theta)$ – called the **prior distribution** – that expresses our beliefs about a parameter θ before we see any data.

2. We choose a statistical model $f(x \mid \theta)$ that reflects our beliefs about x given θ. Notice that we now write $f(x \mid \theta)$ instead of $f(x; \theta)$

3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the **posterior** distribution $f(\theta \mid X_1, \ldots, X_n)$.

The Bayesian Method

Bayes theorem tells us that for discrete variable and discrete parameter,

$$\mathbb{P}(\Theta = \theta \,|\, X = x) = \frac{\mathbb{P}(X = x, \Theta = \theta)}{\mathbb{P}(X = x)} = \frac{\mathbb{P}(X = x \,|\, \Theta = \theta)\mathbb{P}(\Theta = \theta)}{\sum_\theta \mathbb{P}(X = x \,|\, \Theta = \theta)\mathbb{P}(\Theta = \theta)}$$

And for continuous variables we have $f(\theta \,|\, x) = \dfrac{f(x \,|\, \theta)\, f(\theta)}{\int f(x \,|\, \theta)\, f(\theta)\, d\theta}$

The Bayesian Method

If we have n IID observations $X_1, \ldots, X_n$, we replace $f(x \,|\, \theta)$ with

$$f(x_1, \ldots, x_n \,|\, \theta) = \prod_{i=1}^{n} f(x_i \,|\, \theta) = \mathscr{L}_n(\theta).$$

Notation: We will write X^n to mean $(X_1, \ldots, X_n)$ and x^n to mean $(x_1, \ldots, x_n)$. Now,

$$f(\theta \,|\, x^n) = \frac{f(x^n \,|\, \theta) f(\theta)}{\int f(x^n \,|\, \theta) f(\theta)\, d\theta} = \frac{\mathscr{L}_n(\theta) f(\theta)}{c_n} \propto \mathscr{L}_n(\theta) f(\theta)$$

where $c_n = \displaystyle\int \mathscr{L}_n(\theta) f(\theta)\, d\theta$ is the **normalizing constant**, that does not depend on θ.

The Bayesian Method

So the **Posterior is proportional to Likelihood times Prior**:

$f(\theta \mid x^n) \propto \mathscr{L}(\theta) f(\theta)$ Why did we throw away the constant c_n?

Because we can always recover it later if needed.

What do we do with the posterior? First, we get a point estimate – typically, mean $\overline{\theta}_n = \displaystyle\int \theta f(\theta \mid x^n)\, d\theta$ or mode.

We can also obtain a (Bayesian) interval estimate – such (a, b) that $\mathbb{P}(\theta \in (a, b) \mid x^n) = 1 - \alpha$ – by picking a, b such that:

$$\int_{-\infty}^{a} f(\theta \mid x^n)\, d\theta = \int_{b}^{\infty} f(\theta \mid x^n)\, d\theta = \alpha/2$$

The Bayesian Method

Example: Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, where σ is known. Suppose we take as a prior $\theta \sim \mathcal{N}(a, b^2)$. One can then show that the posterior is $\theta \,|\, X^n \sim \mathcal{N}(\overline{\theta} \,|\, \tau^2)$, where $\overline{\theta} = w\overline{X} + (1-w)a$

and $w = \dfrac{1/\text{se}^2}{1/\text{se}^2 + 1/b^2}$, $1/\tau^2 = 1/\text{se}^2 + 1/b^2$, and $\text{se} = \sigma/\sqrt{n}$ is the s.e. of the MLE $\overline{X}$.

Note that $w \to 1$ and $\tau/\text{se} \to 1$ with $n \to \infty$. So, for large n, the posterior is approx. $\mathcal{N}(\widehat{\theta}, \text{se}^2)$. Same is true if n is fixed, but $b \to \infty$ – the prior being very flat.

One can easily find the confidence intervals for $\theta \,|\, X^n \sim \mathcal{N}(\overline{\theta} \,|\, \tau^2)$.

Functions of parameters

Functions of parameters

How do we make predictions about a function $\tau = g(\theta)$? We solved this problem before, posterior CDF is

$$H(\tau \mid x^n) = \mathbb{P}(g(\theta) \leq \tau \mid x^n) = \int_{\theta:\, g(\theta)\leq\tau} f(\theta \mid x^n)\, d\theta$$

so the PDF is $h(\tau \mid x^n) = H'(\tau \mid x^n)$.

Exercise: Let $X_1, \ldots, X_n \sim$ Bernoulli(p) and $f(p) = 1$. So that $p \mid X^n \sim$ Beta$(s+1, n-s+1)$ with $s = \sum_{i=1}^{n} x_i$. Find the CDF and the PDF for $\psi = \log \dfrac{p}{1-p}$.

Simulation

Functions of parameters

A posterior can often be approximated by simulation. Suppose we draw $\theta_1, \ldots, \theta_B \sim p(\theta \mid x^n)$. Then a histogram of $\theta_1, \ldots, \theta_B$ approximates the posterior density $p(\theta \mid x^n)$. Posterior mean $\overline{\theta}_n = \mathbb{E}(\theta \mid x^n)$ is approximated by $B^{-1} \sum_{j=1}^{B} \theta_j$. The posterior $(1 - \alpha)$-interval can be approximated by sample quantiles. For a function of a parameter, $\tau = g(\theta)$, taking $\tau_i = g(\theta_i)$ of $\theta_1, \ldots, \theta_B$ gives a proper sample.

Exercise: Recall the previous slide – looking for posterior of $\psi = \log(p/(1-p))$. 1) draw $P_1, \ldots, P_B \sim$ Beta$(s+1, n-s+1)$, then 2) $\psi_i = \log(P_i/(1-P_i))$ – are IID draws from $h(\psi \mid x^n)$, their histogram provides an estimate of the PDF.

Large sample properties of Bayesian methods

Large sample properties

We saw that for Normal, posterior mean is close to MLE. In general:

Theorem: Let $\widehat{\theta}_n$ be the MLE and $\hat{\text{se}} = 1/\sqrt{nI(\widehat{\theta}_n)}$. Under appropriate regularity conditions, the posterior is approx. Normal with mean $\widehat{\theta}_n$ and std $\hat{\text{se}}$. Hence, $\overline{\theta}_n \approx \widehat{\theta}_n$.
Also, if $C_n = \left(\widehat{\theta}_n - z_{\alpha/2}\,\hat{\text{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\hat{\text{se}}\right)$ is the asymptotic frequentist $(1-\alpha)$-confidence interval, then C_n is also an approximate Bayesian $(1-\alpha)$-confidence interval:
$\mathbb{P}(\theta \in C_n \,|\, X^n) \to 1 - \alpha$

There's also Bayesian delta-method for $\tau = g(\theta)$
$\tau \,|\, X^n \approx \mathcal{N}(\widehat{\tau}, \tilde{\text{se}}^2)$ where $\widehat{\tau} = g(\widehat{\theta})$ and $\tilde{\text{se}} = \hat{\text{se}}\,|\,g'(\widehat{\theta})\,|$.

Flat, Improper and “Noninformative” priors

Flat, Improper and “Noninformative” priors

Where does one get the prior $f(\theta)$? **Subjectivist** would say it should reflect our subjective opinion about θ (before the data is collected). This is possible, but impractical in complicated problems with many parameters. Also, contrary to making scientific inference as objective as possible.

An alternative is **“noninformative”** prior - flat prior $f(\theta) \propto$ const. For Bernoulli, $f(p) = 1$ leads to $p \mid X^n \sim$ Beta$(s+1, n-s+1)$, which is reasonable.

Flat, Improper and “Noninformative” priors

Improper priors. Let $X \sim \mathcal{N}(\theta, \sigma)$ with σ known. If we take a flat prior $f(\theta) \propto c, (c > 0)$ – not a proper probability, since $\int f(\theta) = \infty$.

But we can still use Bayes’ theorem, $f(\theta) \propto \mathcal{L}_n(\theta) f(\theta) \propto \mathcal{L}_n(\theta)$. This gives $\theta \,|\, X^n \sim \mathcal{N}(\overline{X}, \sigma^2/n)$, and the point estimate are exactly frequentist ones. So, ***improper priors are not a problem*** – as long as ***posterior is a proper probability distribution***!

Flat, Improper and “Noninformative” priors

Flat priors are not invariant. Let $X \sim$ Bernoulli(p) and $f(p) = 1$. Let $\psi = \log(p/(1-p))$ – for which $f_\Psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2}$ – is not flat, a contradiction: if ignorant about p, we should also be about ψ

Jeffreys’ priors. Harold Jeffreys came up with a solution to this:

$f(\theta) \propto \sqrt{I(\theta)}$ ($I(\theta)$ is Fisher info) This is transformation invariant!

(In the multiparameter case, $\sqrt{\det I(\theta)}$, $I(\theta)$ being a matrix)

Example: Consider Bernoulli(p). Recall $I(p) = \frac{1}{p(1-p)}$. So Jeffreys’ prior is $f(p) \propto \sqrt{I(p)} = 1/\sqrt{p(1-p)}$. This is Beta$(1/2,1/2)$ – which is very close to a uniform density.