Lecture 4: Bayesian Analysis

Chris Conlon

NYU Stern

February 25, 2019

Quick Refresh: Bayes Rule

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

Given a positive test result what is the probability a patient actually has cancer?

Table: Test Accuracy

	Cancer (1%)	No Cancer (99%)
Positive Test	80%	9.6%
Negative Test	20%	90.4%

Caclulate $Pr(Cancer\&PositiveTest)$ and $Pr(NoCancer\&PositiveTest)$

Table: Joint Probabilities

	Cancer (1%)	No Cancer (99%)
Positive Test	(0.8)(0.01)=0.008	(0.9)(0.096)=0.09504
Negative Test	(0.2)(0.01)=0.002	(0.9)(0.904)=0.89496

$$Pr(Cancer|PositiveTest) = \frac{Pr(Cancer, PosTest)}{Pr(Cancer, PosTest) + Pr(NoCancer, PosTest)} = .008/.10304 = 0.0776$$

- Suppose that we toss a coin several times with $x_i \in \{H, T\} = \{1, 0\}$
- $\mathbf{X} = \{H, T, H, H, \ldots\}$.
- Suppose that the probability of heads $Pr(x_i = H) = p$.
- What is the likelihood of an observed sequence of $\mathbf{X}$? where x_i are I.I.D.

$$Pr(x_i|p) = p^{x_i}(1-p)^{1-x_i}$$
$$Pr(\mathbf{X}|p) = p^{\sum_i x_i}(1-p)^{\sum_i (1-x_i)}$$

Can construct the log likelihood and find the MLE.

$$\ell(\mathbf{X}|p) = (\sum_i x_i) \ln p + (N - \sum_i x_i) \ln(1-p)$$

$$\frac{\partial \ell(p)}{\partial p} = (\sum_i x_i) \frac{1}{p} - (N - \sum_i x_i) \frac{1}{1-p} = 0$$

$$\frac{1-p}{p} = \frac{(\frac{1}{N} \cdot N - \frac{1}{N} \cdot \sum_i x_i)}{\frac{1}{N} \cdot \sum_i x_i} \to \hat{p} = \frac{1}{N} \cdot \sum_i x_i$$

Can also construct the properties of $\hat{p}$.

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{N} \cdot \sum_i x_i\right] = \left[\frac{1}{N} \cdot \sum_i \mathbb{E} x_i\right] = \mu_x = p_0$$
$$\mathbb{V}[\hat{p}|\mathbf{X}] = \mathbb{V}\left[\frac{1}{N} \cdot \sum_i x_i\right] = \frac{1}{N^2} \cdot \sum_i \mathbb{V}(x_i) = \frac{N}{N^2} p(1-p)$$

Which gives us a CI of: $\left(\bar{x} \pm 1.96 \cdot \sqrt{\frac{1}{N}\bar{x}(1-\bar{x})}\right)$

Bayesian Statistics: Brief Introduction

A different idea:

- Start with a (diffuse) initial guess for the distribution of p: $f_P(p)$.
- Incorporate information from likelihood: $f(x_i|p)$
- Construct posterior density estimate $f(p|x_i)$.
  - This doesn't characterize a best estimate $\hat{p}$ but a full distribution.
  - We can calculate $\mathbb{E}[p|x_i]$ or $\mathbb{V}[p|x_i]$ or any other functions of the posterior density.
- Challenge: How to choose initial $f_P(p)$.

One possible guess is the uniform distribution $f(x) = 0$ on $0 \leq x \leq 1$.

- Marginal/Prior Distribution: $f_P(p) = 1$ for $0 \leq p \leq 1$.
- Conditional Distribution/Likelihood: $f_{X|P}(x|p) = p^x(1-p)^{1-p}$
- Joint Distribution :
  $f_{X,P}(x,p) = f_{X|P}(x|p) \cdot f_P(p) = p^x(1-p)^{1-p} \cdot 1 = x \cdot p + (1-x) \cdot (1-p)$
  - This is only defined for $p \in [0,1]$ and $x \in \{0,1\}$. It is zero elsewhere.
- What about Marginal Distribution for x?

$$\int_p f_{PX}(x,p)dp = x \cdot \int_0^1 pdp + (1-x) \cdot \int_0^1 (1-p)dp$$
$$= x \cdot \frac{1}{2} + (1-x) \cdot \frac{1}{2} = \frac{1}{2} \propto 1$$

The object we are usually interested in is the Posterior Distribution

$$f_{P|X}(p|x) = \frac{f_{X|P}(x|p) \cdot f_P(p)}{\int_0^1 f_{X|P}(x|p) \cdot f_P(p)dp} = 2p^x(1-p)^{1-x} \propto p^x(1-p)^{1-x}$$

- We are back at the p.m.f. of the Bernoulli which is maybe comforting.
- This is true because $f_X(x) \propto 1$ and $f_P(p) \propto 1$.
- $f_{P|X}(p|x=0) = (1-p)$ and $f_{P|X}(p|x=1) = p$.

- Let's try a different prior distribution than the uniform we used last time. This time we will use a *Beta*(α, β) distribution:

$$f_P(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\mathbb{E}[p|\alpha,\beta] = \frac{\alpha}{\alpha+\beta}$$

$$\mathbb{V}[p|\alpha,\beta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- This has the advantage that it places nicely with the Binomial.
- Consider $\alpha = 16, \beta = 8$. This gives $\mathbb{E}[p] = \frac{2}{3}$ and $\mathbb{SE}[p] = 0.094$.

Consider the case where $x = 1$ (we get one piece of new data).

$$f_P(p) \cdot f_{X|P}(x|p) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}}_{c(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \cdot p \propto p^{\alpha}(1-p)^{\beta-1}$$

- The resulting distribution is now $(p|x=1) \sim Beta(\alpha+1, \beta)$.
- Our posterior has mean = 0.68 and SE = 0.091.
- Estimate of mean increases and SE decreases.
- Likewise if $x = 0$ we get $(p|x=0) \sim Beta(\alpha, \beta+1)$
- There is a conjugacy relationship between the Beta and the Binomial.

General Case

$$\overbrace{f_{\theta|X}(\theta|x)}^{\text{posterior}} = \frac{\overbrace{f_{X,\theta}(x,\theta)}^{\text{joint}}}{\underbrace{f_X(x)}_{\text{marginal of } x}} = \frac{\overbrace{f_{X|\theta}(x|\theta)}^{\text{likelihood}} \cdot \overbrace{f_\theta(\theta)}^{\text{prior}}}{\int f_{X|\theta}(x|\theta) \cdot f_\theta(\theta) d\theta}$$

There is a shortcut because the denominator doesn't depend on θ

$$f_{\theta|X}(\theta|x) \propto f_{X|\theta}(x|\theta) \cdot f_\theta(\theta) = \mathcal{L}(\theta|x) \cdot f_\theta(\theta)$$

We can cheat because there exists a constant c so that $c \int \mathcal{L}(\theta|x) \cdot f_\theta(\theta) d\theta = 1$.

A Normal Example

Assume $X \sim N(\mu, 1)$ and $\mu \sim N(0, 100)$. What is $f_{\mu|X}(\mu|X = x)$?

$$
\begin{aligned}
f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2 \cdot 100}\mu^2\right) \\
&= \exp -\frac{1}{2}(x^2 - 2x\mu + \mu^2 + \mu^2/100) \\
&\propto \exp\left(-\frac{1}{2(100/101)}(\mu - (100/101)x)^2\right)
\end{aligned}
$$

It happens that $(u|x) \sim N(100x/101, 100/101)$.
In general the posterior will not be well defined.

Kalman Update: A More Complicated Normal

Assume $X \sim N(\mu, \sigma^2)$ with σ^2 known. $\mu \sim N(\mu_0, \tau^2)$. What is $f_{\mu|X}(\mu|X = x)$?

$$\begin{aligned}
f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2\cdot\tau^2}\left(\mu-\mu_0\right)^2\right) \\
&\propto \exp -\frac{1}{2}\left(\frac{x^2}{\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2} + \frac{\mu^2}{\tau^2} - \frac{2\mu\mu_0}{\tau^2} + \frac{\mu_0^2}{\tau^2}\right) \\
&\propto \exp -\frac{1}{2}\left(\mu^2\frac{\sigma^2+\tau^2}{\tau^2\sigma^2} - \mu\frac{2x\tau^2 + 2\mu_0\sigma^2}{\tau^2\cdot\sigma^2}\right) \\
&\propto \exp -\frac{1}{2\left(1/\left(1/\tau^2 + 1/\sigma^2\right)\right)}\left(\left(\mu - \left(x/\sigma^2 + \mu_0/\tau^2\right)/\left(1/\sigma^2 + 1/\tau^2\right)\right)\right.
\end{aligned}$$

The resulting distribution is Normal with mean and variance

$$\mathbb{E}[\mu|X = x] = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \qquad \frac{1}{\mathbb{V}(\mu|X)} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$$

Kalman Update: A More Complicated Normal

Despite being a giant mess this makes sense:

$$\mathbb{E}[\mu|X = x] = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \qquad \frac{1}{\mathbb{V}(\mu|X)} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$$

- Posterior mean is a weighted average of prior mean and sample mean.
- Weights depend on precision of two samples.
- Posterior Precision is sum of precision of each sample $\frac{1}{\mathbb{V}(\cdot)}$
- Probably we want to choose a relatively uninformative prior with large τ^2.
- $\tau^2 \to \infty$ implies an improper prior distribution because it no longer integrates to one. But because of $\propto$ still mostly ok.

Generalization to Multiple Observations

This is straighforward:

$$p(\theta|X_1,\ldots,X_N) \propto \mathcal{L}(\theta|X_1,\ldots,X_N)\cdot p(\theta)$$

- Still depends on: prior, likelihood to construct posterior.
- Can update one observation at a time or all at once.

Bernstein von-Mises Theorem

A posterior distribution converges as you get more and more data to a multivariate normal distribution centred at the maximum likelihood estimator with covariance matrix given by $n^{-1}I(\theta_0)^{-1}$, where θ_0 is the true population parameter (Edit: here $I(\theta_0)$ is the Fisher information matrix at the true population parameter value).

Under these conditions (and some more):

1. MLE is consistent
2. Fixed number of parameters
3. θ_0 in interior of Θ (true value of SD can't = 0).
4. The prior density must be non-zero in a neighborhood of θ_0.
5. log-likelihood needs to be smooth (two derivates at the true value and more)

Conjugate Priors

What is a Conjugate Prior?

For a given likelihood $f_{X|\theta}(x|\theta)$ we can chose a prior $f_\theta(\theta)$ so that the posterior is proportional to a known parametric distribution.

- This makes life easy because now the posterior has a known parametric distribution (normal, beta, gamma, etc.)
- Other than convenience, this alone doesn't tell us that our choice of $f_\theta(\theta)$ is the best prior by any metric.
- Using a non-conjugate prior is entirely defensible, just less convenient.

Empirical Bayes

What is Emprical Bayes?

- Priors can be an important modeling choice
- But what makes a good prior?
  - Sufficiently diffuse
  - As non-informative as possible
  - Don't tip the scales
  - Don't rule out the truth
- Idea: can we use the data itself to construct a prior?
  - If everything is a function of data, are we back in frequentist paradigm?
  - Can we get benefits of Bayes estimation without unpalatable assumptions?

A (famous) Baseball Example

Suppose we want to estimate batting averages (AVG) for some baseball players

- $AVG = \frac{\#\text{hits}}{\#AtBats}$
- Use data on the first $n = 45$ at bats and hits x_i for the 1970 season.
- Predict the batting average μ_i for the end of the season ($n = 400 - 500$ at bats).
- Obvious estimate is batting average after 45 at bats: $\widehat{\mu}_i^{MLE} = x_i/45$.
- Is there a better estimate?

A Baseball Example

Table 1.1: Batting averages $z_i = \hat{\mu}_i^{(\mathrm{MLE})}$ for 18 major league players early in the 1970 season; μ_i values are averages over the remainder of the season. The James–Stein estimates $\hat{\mu}_i^{(\mathrm{JS})}$ (1.35) based on the z_i values provide much more accurate overall predictions for the μ_i values. (By coincidence, $\hat{\mu}_i$ and μ_i both average 0.265; the average of $\hat{\mu}_i^{(\mathrm{JS})}$ must equal that of $\hat{\mu}_i^{(\mathrm{MLE})}$.)

Name	hits/AB	$\hat{\mu}_i^{(\mathrm{MLE})}$	μ_i	$\hat{\mu}_i^{(\mathrm{JS})}$
Clemente	18/45	.400	**.346**	.294
F Robinson	17/45	.378	**.298**	.289
F Howard	16/45	.356	**.276**	.285
Johnstone	15/45	.333	**.222**	.280
Berry	14/45	.311	**.273**	.275
Spencer	14/45	.311	**.270**	.275
Kessinger	13/45	.289	**.263**	.270
L Alvarado	12/45	.267	**.210**	.266
Santo	11/45	.244	**.269**	.261
Swoboda	11/45	.244	**.230**	.261
Unser	10/45	.222	**.264**	.256
Williams	10/45	.222	**.256**	.256
Scott	10/45	.222	**.303**	.256
Petrocelli	10/45	.222	**.264**	.256
E Rodriguez	10/45	.222	**.226**	.256
Campaneris	9/45	.200	**.286**	.252
Munson	8/45	.178	**.316**	.247
Alvis	7/45	.156	**.200**	.242
Grand Average		.265	**.265**	.265

A (famous) Baseball Example

Probably we can do better than the MLE here:

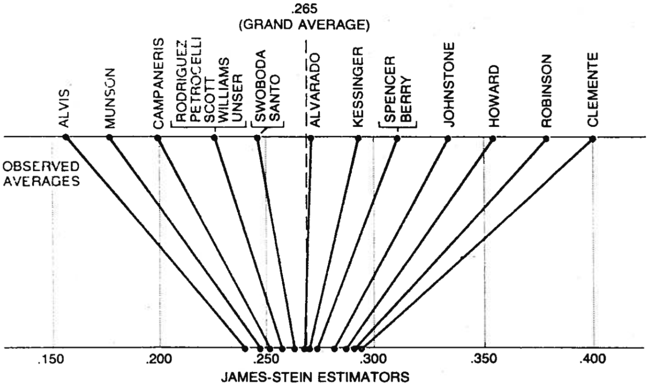
- Thurman Munson wins Rookie of the Year and ends up batting $\mu_i = .316$. If he batted .178 all year, his career would not have lasted long.
- Clemente's .400 seems unlikely to hold up. Last player to hit $>$ .400 was Ted Williams .406 in 1941.
- But how?

Idea is to take an average between the observed average y_i and the overall mean $\bar{y}$:

$$\widehat{\mu}_i^{JS} = (1 - \lambda) \cdot \bar{y} + \lambda \cdot y_i, \quad \lambda = 1 - \frac{(m-3)\sigma^2}{\sum_i (y_i - \bar{y})^2}$$

- This has the effect of shrinking y_i towards the prior mean $\bar{y}$.
- In this case the prior mean is just $\bar{y}$ the grand-mean of all players
- How can information about unrelated players inform us about μ_i?
- Also consider proportion of foreign cars in Chicago as an additional y_i, can this help too?
- The shrinkage factor λ depends on sample size and variance, but how is it chosen?

A Baseball Example



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.