WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

CS639:
Data Management for Data Science

Lecture 14: Bayesian Methods

Theodoros Rekatsinas

Announcements

- We will release grades of Midterm by the end of day today.

Today – Bayesian Methods

- Motivation and Introduction
- Bayes Theorem
- Bayesian inference

Motivation

- Statistical inference: Drawing conclusions based on data that is subject to random variation (observational errors and sampling variation)
- So far we saw the “frequentists” point of view.
- Bayesian inference provides a different way to draw conclusions from data.

Basic Idea

- Leverage **prior information** and update prior information with new data to create a **posterior probability distribution**.

- Three steps:
  - Form prior (a probability model)
  - Condition on observed data (new data from your sample)
  - Evaluate the posterior distribution

Basic Idea

- “The **central feature** of Bayesian inference [is] the **direct quantification of uncertainty**” (Gelman et al. 2014, 4).
- Less emphasis on p-value hypothesis testing. More emphasis on the confidence and probability intervals.
- Many researchers actually interpret ‘frequentist’ confidence intervals *as if* they were Bayesian probability intervals.

Uncertainty in Freq. and Bayesian Approaches

- Both involve the **estimation of unknown quantities** of interest
- The estimates they produce have **different interpretations**.

- **Frequentist**: **95% Confidence interval**: Repeated samples will contain the true parameter within the interval 95% of the time.

- **Bayesian**: **95% Probability (credible) interval**: There is a 95% **probability** that the unknown parameter is actually in the interval.

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?

- We denote random variables with capital letters

- Random variables have domains
  - R in {true, false} (sometimes write as {+r, ¬r})
  - D in $[0, \infty)$
  - L in possible locations, maybe {(0,0), (0,1), …}

Probability Distributions

- Discrete random variables have distributions

$$P(T)$$

T	P
warm	0.5
cold	0.5

$$P(W)$$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

- A discrete distribution is a TABLE of probabilities of values
- The probability of a state (lower case) is a single number

$$P(W = rain) = 0.1 \qquad P(rain) = 0.1$$

- Must have:

$$\forall x \; P(x) \geq 0 \qquad \sum_x P(x) = 1$$

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment:

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - How many assignments if *n* variables with domain sizes *d*?
  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- For all but the smallest distributions, impractical to write out or estimate
  - Instead, we make additional assumptions about the distribution

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_w P(t,w)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(w) = \sum_t P(t,w)$$

$P(W)$

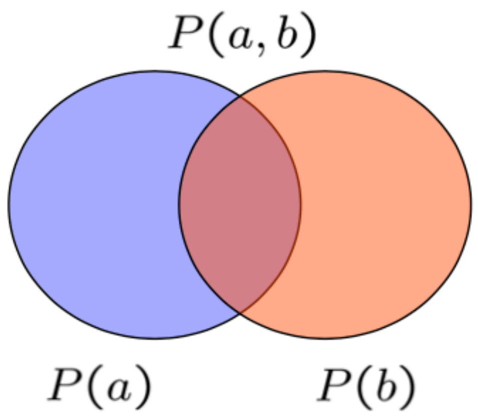
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=r|T=c) = ???$$

Conditional Probabilities

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

The Product Rule

- Sometimes have conditional distributions but want the joint

$$P(x|y) = \frac{P(x,y)}{P(y)} \quad \Longleftrightarrow \quad P(x,y) = P(x|y)P(y)$$

- Example:

$P(W)$

W	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Longleftrightarrow$

$P(D,W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

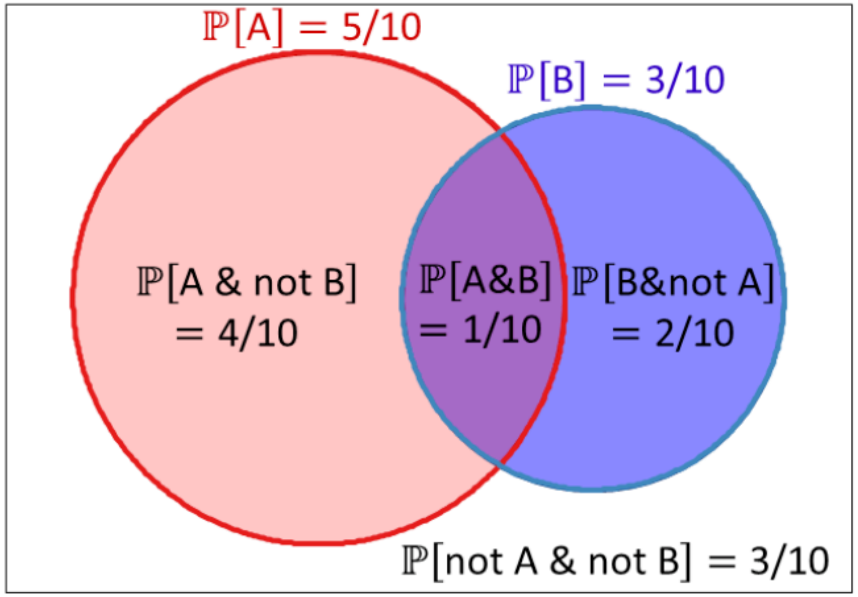
$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Let's us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many practical systems (e.g. ASR, MT)

Bayes' Theorem

Before we get to inference: Bayes' *Theorem* is a result in conditional probability, stating that for two events A and B...

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \text{ and } B]}{\mathbb{P}[B]} = \mathbb{P}[B|A]\frac{\mathbb{P}[A]}{\mathbb{P}[B]}.$$



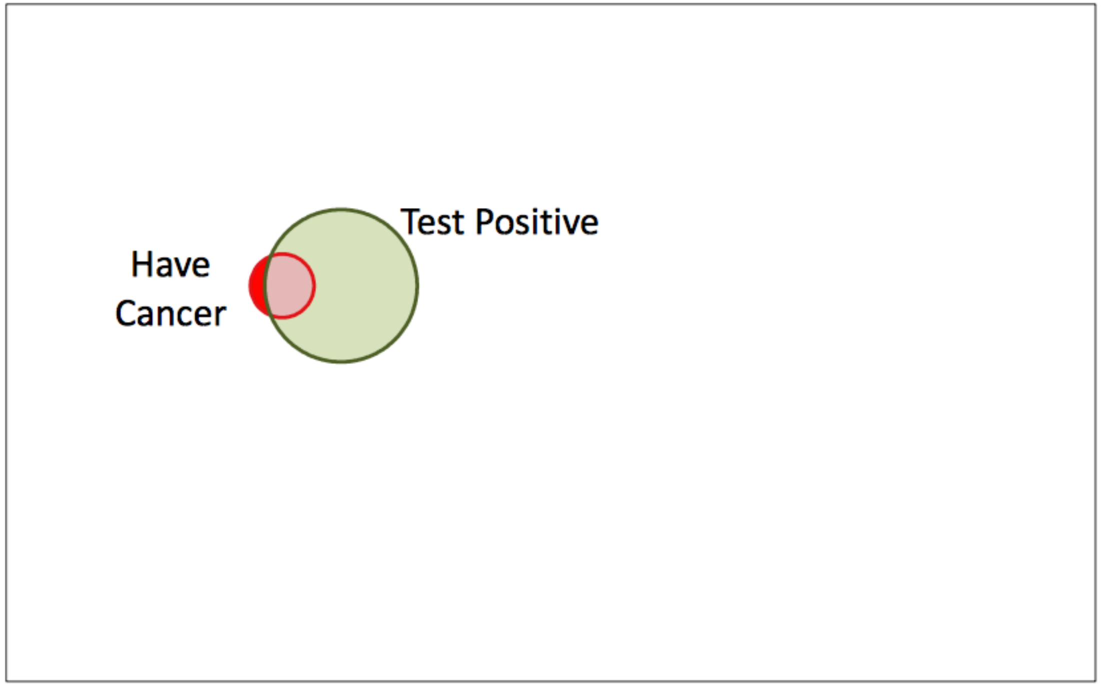
In this example;

- $\mathbb{P}[A|B] = \frac{1/10}{3/10} = 1/3$
- $\mathbb{P}[B|A] = \frac{1/10}{5/10} = 1/5$
- And $1/3 = 1/5 \times \frac{5/10}{3/10}$ ($\checkmark$)

In words: the conditional probability of A given B is the conditional probability of B given A scaled by the *relative* probability of A compared to B.

Bayes' Theorem

Why does it matter? If 1% of a population have cancer, for a screening test with 80% sensitivity and 95% specificity;



$$\begin{aligned} \mathbb{P}[\,\text{Test } {+}\text{ve}|\text{Cancer}\,] &= 80\% \\ \frac{\mathbb{P}[\,\text{Test } {+}\text{ve}]}{\mathbb{P}[\,\text{Cancer}\,]} &= 5.75 \\ \mathbb{P}[\,\text{Cancer}|\text{Test } {+}\text{ve}\,] &\approx 14\% \end{aligned}$$

... i.e. most positive results are actually false alarms

Mixing up $\mathbb{P}[\,A|B\,]$ with $\mathbb{P}[\,B|A\,]$ is the *Prosecutor's Fallacy*; a small probability of evidence given innocence need NOT mean a small probability of innocence given evidence.

Bayesian Approach

How to update knowledge, as data is obtained? We use;

- **Prior distribution:** what you know about parameter $\boldsymbol{\beta}$, excluding the information in the data – denoted $\pi(\boldsymbol{\beta})$
- **Likelihood:** based on modeling assumptions, how [relatively] likely the data $\mathbf{Y}$ are *if* the truth is $\boldsymbol{\beta}$ – denoted $f(\mathbf{Y}|\boldsymbol{\beta})$

So how to get a **posterior distribution:** stating what we know about $\boldsymbol{\beta}$, combining the prior with the data – denoted $p(\boldsymbol{\beta}|\mathbf{Y})$? Bayes Theorem *used for inference* tells us to multiply;

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{Y}) &\propto f(\mathbf{Y}|\boldsymbol{\beta}) \times \pi(\boldsymbol{\beta}) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior.} \end{aligned}$$

... and that's it! (essentially!)

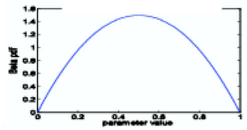
- No replications – e.g. no replicate plane searches
- Given modeling assumptions & prior, process is automatic
- Keep adding data, and updating knowledge, as data becomes available... knowledge will concentrate around true $\boldsymbol{\beta}$

Bayesian Learning

- Use Bayes' rule!

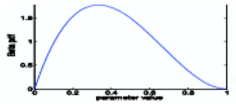
Data Likelihood

Prior



$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Posterior



Normalization

- Or equivalently: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$
- For *uniform* priors, this reduces to maximum likelihood estimation!

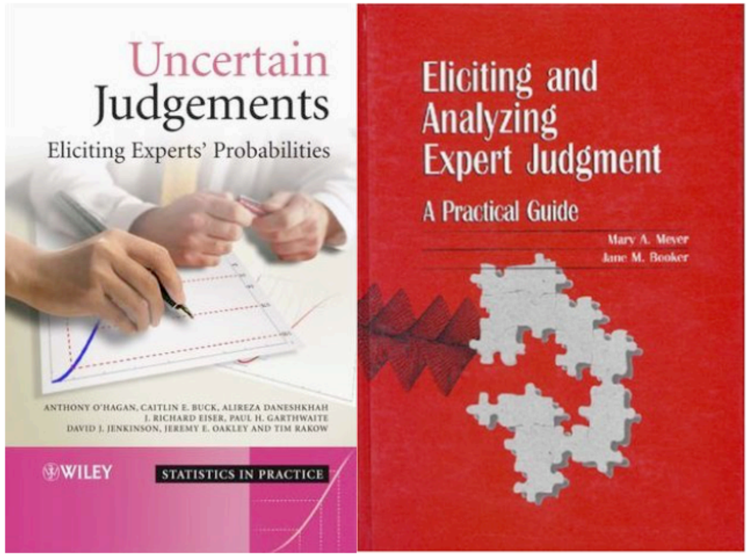
$$P(\theta) \propto 1 \qquad P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)$$

Where do priors come from?

Priors come from all data *external* to the current study, i.e. everything else.

'Boiling down' what subject-matter experts know/think is known as *eliciting* a prior.

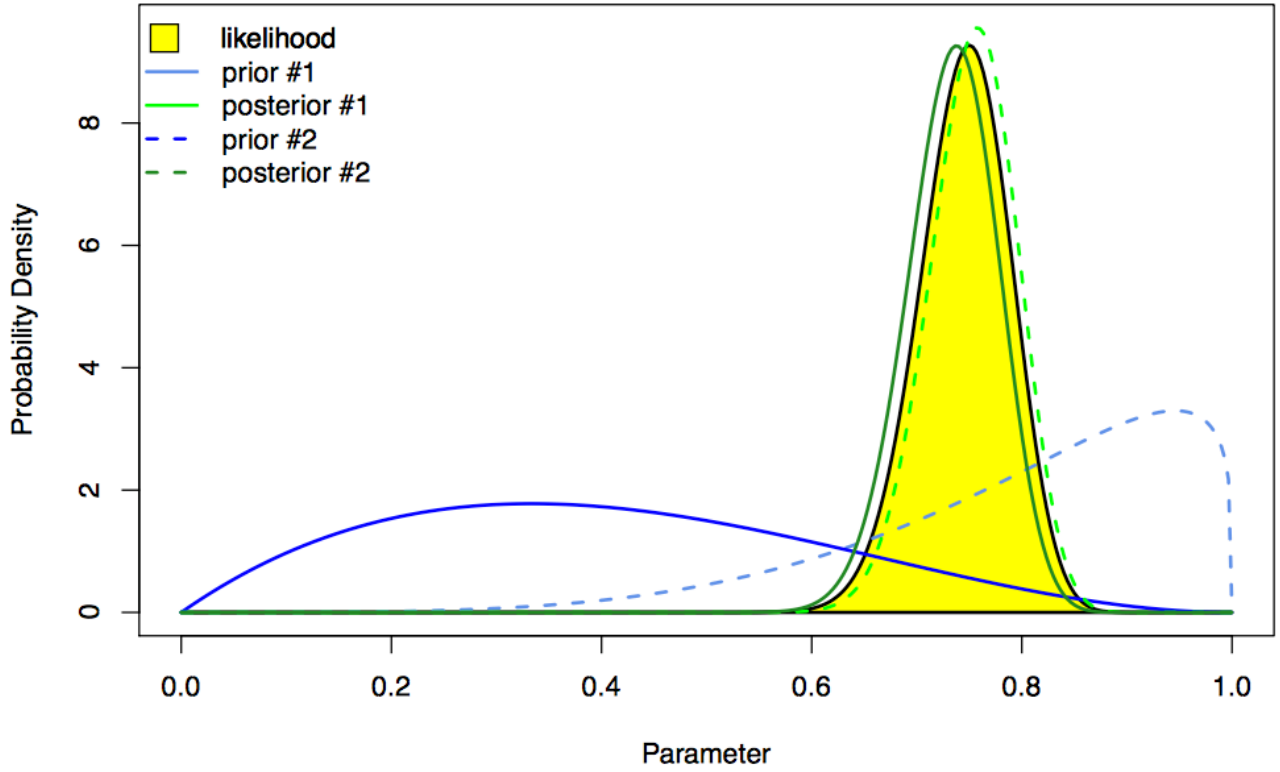
It's not easy (see right) but here are some simple tips;



- Discuss parameters experts understand – e.g. code variables so intercept is mean outcome in people with average covariates, *not* with age=height=IQ=0
- Avoid leading questions (just as in survey design)
- The 'language' of probability is unfamiliar; help users express their uncertainty

When don’t prior matter (much)?

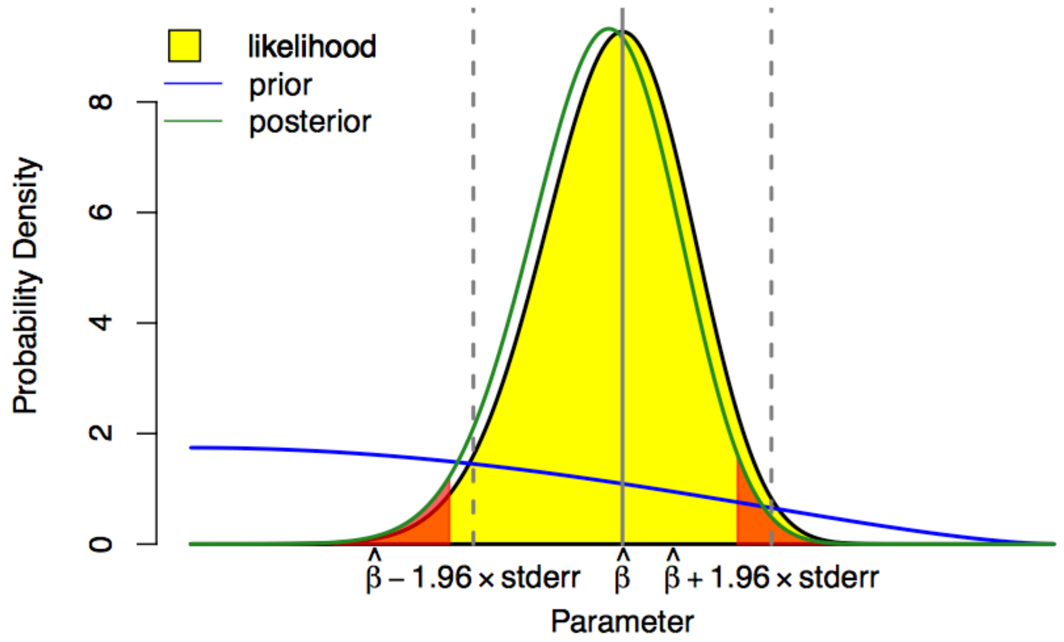
When the data provide a lot more information than the prior, this happens; (recall the stained glass color-scheme)



These priors (& many more) are *dominated* by the likelihood, and they give very similar posteriors – i.e. everyone agrees. (Phew!)

When don’t prior matter (much)?

Back to having very informative data – now zoomed in;



The likelihood *alone* (yellow) gives the classic 95% confidence interval. But, to a good approximation, it goes from 2.5% to 97.5% points of Bayesian posterior (red) – a 95% *credible* interval.

- With large samples*, sane frequentist confidence intervals and sane Bayesian credible intervals are essentially identical
- With large samples*, it's actually *okay* to give Bayesian interpretations to 95% CIs, i.e. to say we have ≈95% posterior belief that the true β lies within that range

* *and some regularity conditions*

Summary

Bayesian statistics:

- Is useful in many settings, and you should know about it
- Is *often* not very different *in practice* from frequentist statistics; it is often helpful to think about analyses from both Bayesian and non-Bayesian points of view
- Is not reserved for hard-core mathematicians, or computer scientists, or philosophers. If you find it helpful, use it.