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
  - $R = \text{Is it raining?}$
  - $D = \text{How long will it take to drive to work?}$
  - $L = \text{Where am I?}$

- We denote random variables with capital letters

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

- Discrete random variables have distributions

\[
P(T) \quad \quad \quad P(W)
\begin{array}{|c|c|}
\hline
T & P \\
\hline
\text{warm} & 0.5 \\
\hline
\text{cold} & 0.5 \\
\hline
\end{array}
\begin{array}{|c|c|}
\hline
W & P \\
\hline
\text{sun} & 0.6 \\
\hline
\text{rain} & 0.1 \\
\hline
\text{fog} & 0.3 \\
\hline
\text{meteor} & 0.0 \\
\hline
\end{array}
\]

- A discrete distribution is a **table** of probabilities of values
- The probability of a state (lower case) is a single number
  \[
P(W = \text{rain}) = 0.1 \quad \quad P(\text{rain}) = 0.1\]
- Must have:
  \[
\forall x \ P(x) \geq 0 \quad \text{and} \quad \sum_{x} P(x) = 1
\]
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:
  $$P(x_1, x_2, \ldots x_n) \geq 0$$
  $$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

- 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

$$P(T, W)$$

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

$$P(T)$$

<table>
<thead>
<tr>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>0.5</td>
</tr>
<tr>
<td>cold</td>
<td>0.5</td>
</tr>
</tbody>
</table>

$$P(W)$$

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.6</td>
</tr>
<tr>
<td>rain</td>
<td>0.4</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\[
P(W = r|T = c) = ????
\]
Conditional Probabilities

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

<table>
<thead>
<tr>
<th>Conditional Distributions</th>
<th>Joint Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(W</td>
<td>T = \text{hot})$</td>
</tr>
<tr>
<td>W</td>
<td>T</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
</tr>
<tr>
<td>sun</td>
<td>sun</td>
</tr>
<tr>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>rain</td>
<td>hot</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>$P(W</td>
<td>T = \text{cold})$</td>
</tr>
<tr>
<td>W</td>
<td>sun</td>
</tr>
<tr>
<td>W</td>
<td>rain</td>
</tr>
<tr>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>rain</td>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>
The Product Rule

- Sometimes have conditional distributions but want the joint

\[ P(x|y) = \frac{P(x, y)}{P(y)} \quad \text{↔} \quad P(x, y) = P(x|y)P(y) \]

- Example:

|       | P(D|W)       | P(D, W)       |
|-------|--------------|---------------|
|       | D  | W  | P  |          | D  | W  | P  |
|       | wet | sun | 0.1 |          | wet | sun | 0.08 |
|       | dry | sun | 0.9 |          | dry | sun | 0.72 |
|       | wet | rain | 0.7 |          | wet | rain | 0.14 |
|       | dry | rain | 0.3 |          | 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 $B$...

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

In this example;

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

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;

\[
\frac{P[\text{Test +ve}|\text{Cancer}]}{P[\text{Cancer}]} = 5.75 \quad \Rightarrow \quad P[\text{Cancer}|\text{Test +ve}] \approx 14% 
\]

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

Mixing up \( P[A|B] \) with \( P[B|A] \) is the \textit{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 $\beta$, excluding the information in the data – denoted $\pi(\beta)$
- **Likelihood**: based on modeling assumptions, how [relatively] likely the data $Y$ are if the truth is $\beta$ – denoted $f(Y|\beta)$

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

$$p(\beta|Y) \propto f(Y|\beta) \times \pi(\beta)$$

Posterior $\propto$ Likelihood $\times$ Prior.

... 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 $\beta$
Bayesian Learning

- Use Bayes’ rule!
  \[ P(\theta | \mathcal{D}) = \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})} \]

- Or equivalently:  
  \[ P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta) \]

- For uniform priors, this reduces to maximum likelihood estimation!
  \[ P(\theta) \propto 1 \quad P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta) \]
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 \( \approx 95\% \) posterior belief that the true \( \beta \) 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.