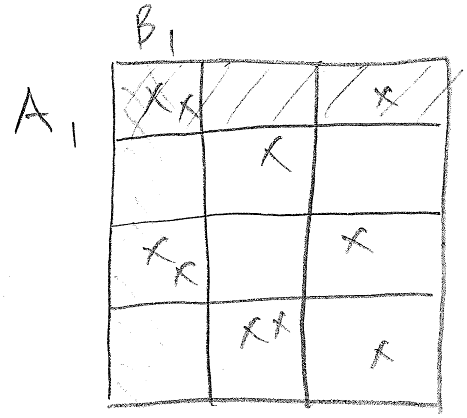
From probability to Bayes rule

Lecture 1

- Bayes rule pops out of basic manipulations of probability distributions.

- Let's reach it through a very simple example.

Example



- Call the entire space Ω
- A_i is the ith row partition
- B_i is the ith column partition

(defined arbitrarily)

- We have points lying in this space, denoted by x's. pick one of these points uniformly at random.

- In this case, calculating probabilities is simply a matter of counting

$$P(x \in A_1) = \frac{\# A_1}{\# \Omega}, \quad P(x \in B_1) = \frac{\# B_1}{\# \Omega}$$

- What about $P(x \in A_1 \mid x \in B_1)$?

This is the probability $x \in A_1$ given that I know $x \in B_1$

This is called a conditional probability.

- Looking at the picture, and doing a clever trick...

$$P(X \in A_1 | X \in B_1) = \frac{\#(A_1 \cap B_1)}{\# B_1}$$

$$= \frac{\#(A_1 \cap B_1)}{\# B_1} \frac{\# \Omega}{\# \Omega} \leftarrow \text{trick}$$

$$= \frac{P(X \in A_1 \;\&\; X \in B_1)}{P(X \in B_1)}$$

- We're just multiplied and divided by the same thing, but already we've made a general statement.

A more general statement

- Let A and B be two events, then

$$P(A|B) = \frac{P(A,B)}{P(B)} \underset{\text{algebra}}{\Longrightarrow} P(A|B)P(B) = P(A,B)$$

We have some names for these:

$P(A|B)$: conditional distribution
$P(A,B)$: joint distribution
$P(B)$: marginal distribution

- This last one is tricky since it's also just "the probability of B." However we can also express it as follows:

Bayes rule

- So we have that $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- These values each have a name:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- 1. Imagine we don't know A, but we have some <u>prior</u> belief about its value
2. We get some information about A in the form of <u>data</u> B.
3. Bayes rule tells us a mathematically rigorous way to incorporate this information in our belief about A in the form of the <u>posterior</u> distribution.

- In reference to the picture,

$$P(B) = \frac{\#B}{\#\Omega} = \frac{\sum_i \#(A_i \cap B)}{\#\Omega} = \sum_i \frac{\#(A_i \cap B)}{\#\Omega} = \sum_i P(A_i, B)$$

bring inside the summation

- Requirements: $A_i \cap A_j = \phi$, $\bigcup_i A_i = \Omega$

Getting to Bayes rule

- We're a few easy steps away

Showed that: $P(A,B) = P(A|B)P(B)$ therefore the

by symmetry: $P(A,B) = P(B|A)P(A)$ RHS are equal

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

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

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

- This is called Bayes rule.

- As you can see, there are a few ways it can be written.

Example (medical test)

- This classic example shows how Bayes rule can lead to counter-intuitive results.

- We have two binary indicators.

$$A_i = \begin{cases} 1 & \text{person } i \text{ has disease} \\ 0 & \text{no disease} \end{cases}$$ ← don't get to observe this

$$B_i = \begin{cases} 1 & \text{test for disease is positive} \\ 0 & \text{test is negative} \end{cases}$$ ← get to observe this

- Person i tests "positive" for the disease ($B_i = 1$). What is the probability the person has it ($A_i = 1$)?

- Mathematical problem: Want $P(A_i = 1 \mid B_i = 1)$.
  Does Bayes rule help?

Bayes rule: $$P(A=1 \mid B=1) = \frac{P(B=1 \mid A=1)P(A=1)}{P(B=1)}$$

$$= \frac{P(B=1 \mid A=1)\,P(A=1)}{P(B=1 \mid A=1)p(A=1) + P(B=1 \mid A=0)P(A=0)}$$

- We estimate from historical data that

$P(B=1 \mid A=1) = 0.95$, $P(B=1 \mid A=0) = 0.05$

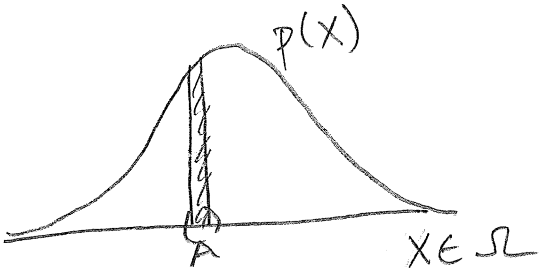
$P(A=1) = 0.01$, $P(A=0) = 0.99$

- Then plugging in, $P(A=1 \mid B=1) = 0.16$

Continuous space

- We've been talking about discrete distributions so far. The number of values the unknowns can take is finite
- When values are in a continuous space, we switch to continuous distributions

Example



observation $X \in \Omega$, $p(X)$ is its <u>density</u>

$$p(X) \geq 0, \quad \int_{\Omega} p(X)\,dx = 1$$

$$p(X \in A) = \int_A p(X)\,dx$$

$$\downarrow$$

$$Pr(X) = p(X)\,dX = 0 \quad \text{(theory!)}$$

- Probability theory lets us ignore probabilities and just work with the densities
- The same rules apply as for discrete random variables.
  - $p(X|\theta) = \dfrac{p(X,\theta)}{p(\theta)}, \quad p(\theta) = \int p(X,\theta)\,dX$
  - $p(X,\theta) = p(X|\theta)p(\theta) = p(\theta|X)p(X)$
- This leads to Bayes rule for continuous variables

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

- The difference is we're dealing with <u>continuous</u> functions

Bayesian modeling

- Applying Bayes rule to the unknown variables of a data modeling problem is called Bayesian modeling.

- In a simple, generic form we can write the model

$$X \sim p(X|\theta) \leftarrow$$ data-generating distribution
this is the model of the data

$$\theta \sim p(\theta) \leftarrow$$ the model prior distribution. Our belief about θ a priori.

- We want to learn θ, so we use Bayes rule

$$\underset{\substack{\downarrow \\ \text{don't know this yet}}}{p(\theta|X)} = \underbrace{p(X|\theta)p(\theta)}_{\text{do know this because we defined it}} / p(X) \leftarrow$$

do we know this? can we calculate $p(X) = \int p(X|\theta)p(\theta)d\theta$?

- This is the general form of the problems discussed in this class. However, it's non-trivial because

1. $p(X|\theta)$ can be quite complex — model definition
2. $p(\theta)$ can be even more complex
3. $p(\theta|X)$ can be <u>intractable</u>. We can't calculate $p(X)$, so we need an algorithm to approximate this.

Simple example: beta-Bernoulli

- We have a sequence of observations $X_1, X_2, \ldots, X_N$ where $X_i = \begin{cases} 1, & \text{"success"} \\ 0, & \text{"failure"} \end{cases}$. Think of them as coin flips.

- We hypothesize that each X_i is generated by flipping a biased coin, $X_i \sim p(X \mid \theta) \Rightarrow p(X_i = 1 \mid \theta) = \theta$.

- We assume the X_i are independent and identically distributed (i.i.d.) according to $p(X \mid \theta) \rightarrow X_i \overset{iid}{\sim} p(X \mid \theta)$.

- This means that the X_i are conditionally independent given the bias θ.

$$p(X_1 \ldots X_N \mid \theta) = \prod_{i=1}^{N} p(X_i \mid \theta).$$

- Since $p(X_i \mid \theta) = \theta^{X_i}(1-\theta)^{1-X_i}$, we can write

$$p(X_1 \ldots X_N \mid \theta) = \prod_{i=1}^{N} \theta^{X_i}(1-\theta)^{1-X_i}$$

- We are interested in the posterior distribution of θ given $X_1 \ldots X_N$, i.e.,

$$p(\theta \mid X_1 \ldots X_N) = p(X_1 \ldots X_N \mid \theta)\, p(\theta) / p(X_1 \ldots X_N)$$

$$= \prod_{i=1}^{N} p(X_i \mid \theta)\, p(\theta) / p(X_1 \ldots X_N)$$

- We've come across our first significant Bayesian problem: what do we set $p(\theta)$ to?

First try

- Let $p(\theta) = \text{Uniform}(0,1) \Rightarrow p(\theta) = \mathbb{1}(0 \le \theta \le 1)$
- Then by Bayes rule and i.i.d. assumption

$$p(\theta \mid X_1 \ldots X_N) = \frac{\prod_{i=1}^N p(X_i \mid \theta)\, p(\theta)}{p(X_1 \ldots X_N)} =$$

$$= \frac{\theta^{\sum_{i=1}^N X_i} (1-\theta)^{N-\sum_{i=1}^N X_i} \mathbb{1}(0 \le \theta \le 1)}{\int_0^1 \theta^{\sum_{i=1}^N X_i} (1-\theta)^{N-\sum_{i=1}^N X_i} \mathbb{1}(0 \le \theta \le 1)\, d\theta}$$

- This normalizing constant is tricky, but fortunately its been solved and we can conclude

$$p(\theta \mid X_1 \ldots X_N) = \frac{\Gamma(N)}{\Gamma(\sum_i X_i)\Gamma(N - \sum_i X_i)} \theta^{\sum_{i=1}^N X_i + 1 - 1} (1-\theta)^{N - \sum_{i=1}^N X_i + 1 - 1}$$

$\Gamma(\cdot)$ is a "gamma function"

- This is a very common distribution called a <u>beta distribution</u>

$$\text{Beta}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

- In the posterior, $a = 1 + \sum_{i=1}^N X_i$, $b = 1 + N - \sum_{i=1}^N X_i$
- Notice that when $a = b = 1$, $\text{Beta}(1,1) = \text{Uniform}(0,1)$ which was our prior.

A "Conjugate" prior

- The beta distribution $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$ looks a lot like the likelihood term, $\theta^{\sum_i x_i} (1-\theta)^{N-\sum_i x_i}$.

- Also, because Unif(0,1) is a special case, a beta prior would give us more options to express belief.

- Beta prior:

"proportional to"

$$p(\theta \mid x_1 \dots x_N) \propto p(x_1 \dots x_N \mid \theta)\, p(\theta)$$

$$\propto \left[\theta^{\sum_i x_i} (1-\theta)^{N-\sum_i x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}\right]$$

$$\propto \theta^{a+\sum_{i=1}^N x_i - 1} (1-\theta)^{b+N-\sum_{i=1}^N x_i - 1}$$

Comments

1. Where did $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ go? We are writing the posterior as a proportionality. The functions

$g(\theta) \propto f(\theta)$ if $g(\theta) = \frac{1}{Z} f(\theta)$ for some constant Z (i.e., Z is not a function of θ)
but may be function of other things

2. In general, if $g(\theta) = \frac{f(\theta)}{\int f(\theta) d\theta}$, then $g(\theta) = \frac{\frac{1}{Z} f(\theta)}{\int \frac{1}{Z} f(\theta) d\theta}$

In both cases we can write $g(\theta) \propto f(\theta) \propto \frac{1}{Z} f(\theta)$

3. In general, we can multiply the joint likelihood by any constant to make it look nicer. In this case we chose $\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

- So we have $p(\theta \mid X_1 \ldots X_N) \propto \theta^{a+\sum_{i=1}^{N} X_i - 1}(1-\theta)^{b+N-\sum_{i=1}^{N} X_i - 1}$

Solution : $$P(\theta \mid X_1 \ldots X_N) = \frac{\theta^{a+\sum_i X_i - 1}(1-\theta)^{b+N-\sum_i X_i - 1}}{\int_0^1 \theta^{a+\sum_i X_i - 1}(1-\theta)^{b+N-\sum_i X_i - 1}\, d\theta}$$

Trick : Notice $\theta^{a+\sum_{i=1}^{N} X_i - 1}(1-\theta)^{b+N-\sum_{i=1}^{N} X_i - 1} \propto \text{Beta}(a', b')$

where $a' = a + \sum_{i=1}^{N} X_i$, $b' = b + N - \sum_{i=1}^{N} X_i$

- The posterior distribution is in the same family as the prior (beta). We just update its parameters.

- Conjugate priors: Let $X \sim p(X \mid \theta)$ and $\theta \sim p(\theta)$. If the posterior $p(\theta \mid X)$ is in the same family as the prior, $p(\theta)$, then $P(\theta)$ is conjugate to the likelihood $p(X \mid \theta)$.

~~Example: gamma - Poisson~~

What do we gain by being Bayesian?

- Consider the expectation an variance under the posterior

$$\mathbb{E}[\theta] = \int_0^1 \theta p(\theta|\vec{X})\,d\theta = \frac{a + \sum_{i=1}^{N} X_i}{a+b+N}$$

$$Var(\theta) = \int_0^1 (\theta - \mathbb{E}[\theta])^2 p(\theta|\vec{X})\,d\theta = \frac{(a + \sum_{i=1}^{N} X_i)(b + N - \sum_{i=1}^{N} X_i)}{(a+b+N)^2(a+b+N-1)}$$

- As N increases,
  1. $\mathbb{E}(\theta) \to$ empirical success rate
  2. $Var(\theta) \to 0$ at the rate $1/N$

- Compare with maximum likelihood

$$\theta_{ML} = \arg\max_{\theta} P(X_1 \ldots X_N | \theta) = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- The Bayesian approach is capturing our uncertainty about the quantity we are interested in gauging
- Maximum Likelihood doesn't do this

- There are more compelling reasons in the case of data modeling for machine learning applications that we will discuss throughout the semester.