Midterm Review

DS6040 Fall 2024
Teague R. Henry

SCHOOL of DATA SCIENCE

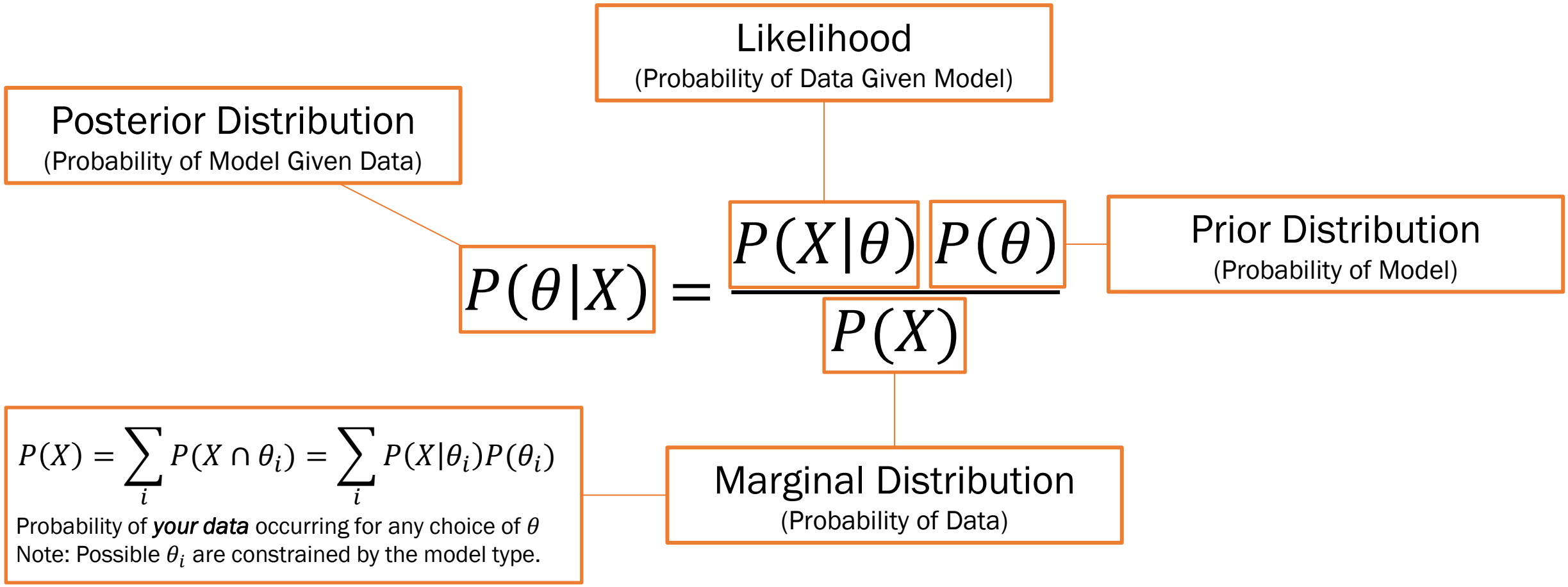
1

Outline

- Bayes Theorem
- Probability Rules
- Priors
- Samplers

2

Bayes' Theorem



Bayes Theorem Components

Likelihood $P(X|\theta)$:

- Probability of the data given the model/parameters.
- This is your model, you specify it.

Priors $P(\theta)$:

- Prior probability of your parameters.
- Represents what you think your parameter values are going to be.
  - Could be based on previous data, could be based on theory, could be revealed to you in a dream.
- You specify this.
- You need a prior distribution for each parameter in your model.

4

Bayes Theorem Components

Marginal $P(X)$:

- Probability of your data across all possible values of your parameters.
- Think of this as a weighted average of probabilities with respect to the priors.
  - Parameter values that are unlikely under your priors won't add anything to the marginal.
- You do not specify this. Estimating this/getting rid of this is the whole purpose of samplers/variational inference/conjugacy.
  - In very simple examples, you do calculate this. For example, in the medical testing scenario.

Posterior $P(\theta|X)$:

- Probability of your parameters given your data (and priors).
- This is the goal of Bayesian inference, these are the results.
- A posterior is always a balance between the data and the priors.

5

Thinking with Probabilities

To solve problems in probability, it helps to think about conditional probabilities.

- Conditioning lets you fix the values of one thing, which will simply the probabilities of the other things in the problem.

Law of total probability:

- Consider a problem where we have 2 groups, A and B, and are interested in the probability of X across both groups.
- We might have $P(A), P(B), P(X|A), P(X|B)$. How do we use these to determine the $P(X)$?
- $P(X) = P(X \cap A) + P(X \cap B)$ – The probability of X is the probability of X and A, plus X and B.
- $P(X) = P(X|A)P(A) + P(X|B)P(B)$ - By the identity of joint probabilities to conditionals.

6

Thinking with Probabilities

The most difficult thing to keep track of in probabilities is when the problem asks you to flip a conditional.

- What is the probability of A given a value of X.

To do this, we must use Bayes Theorem.

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

Write your conditional probabilities out in notation, that will help you know when to use Bayes Theorem.

7

A Problem

An Urn contains 3 decks of cards:

- Deck A: A standard 52 card deck
- Deck B: A deck with only face cards from all four suits (12 cards in total).
- Deck C: A deck consisting of 4 copies of each suits ace (16 cards total).

You perform the following experiment:

1. Randomly select one deck from the urn.
2. Shuffle the deck.
3. Draw a card

Question 1: What is the probability you draw an ace?

8

A Problem

Question 1: What is the probability you draw an ace?

- $P(A) = P(B) = P(C) = \frac{1}{3}$
- $P(Ace|A) = \frac{4}{52}, P(Ace|B) = 0, P(Ace|C) = 1$
- $P(Ace) = P(Ace \cap A) + P(Ace \cap B) + P(Ace \cap C)$
- $P(Ace) = P(Ace|A)P(A) + P(Ace|B)P(B) + P(Ace|C)P(C)$
- $P(Ace) = \frac{4}{52} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{4}{156} + \frac{1}{3} = \frac{1}{39} + \frac{13}{39} = \frac{14}{39} =$

9

A Problem

Question 2: If you draw an Ace, what is the probability that you selected Deck C?

- $P(A) = P(B) = P(C) = \frac{1}{3}$
- $P(Ace|A) = \frac{4}{52}, P(Ace|B) = 0, P(Ace|C) = 1$
- $P(C|Ace) = \frac{P(Ace|C)P(C)}{P(Ace)} = \frac{1 \times \frac{1}{3}}{\frac{14}{39}} = \frac{13}{14}$

10

Priors

Conjugate Priors

- A conjugate prior is a prior that, when combined with the likelihood, leads to a posterior that is from the same family of distributions as the prior.
- For example, a beta-binomial model has a binomial likelihood (which has a single parameter p denoting the probability of success), a beta prior on p, and the posterior of $P(p|data)$ is a beta distribution.
- Conjugate priors allow us to jump directly to the posterior distribution without the need to approximating the normalizing constant.
  - We know the posterior distributions parameters. Therefore we know the expected value, credible intervals, etc etc.
- This makes conjugacy very useful for computational purposes.
- But, there are only a very limited number of models that have conjugate priors…

11

Informative vs. Uninformative Priors

The choice of prior hyperparameters determine if the prior is informative or not.

An informative prior has low uncertainty compared to the scale of the parameter.

- For example, if the prior is on a mean, and is normal, then having a relatively low SD results in an informative prior.
- An informative prior pulls the posterior in the direction of the prior more than an uninformative prior.
- No prior is truly uninformative, not even a Jeffreys prior. An “uninformative prior” is still highly informative when there is not much data available.

We define "(un)informative" as how quickly does the data overwhelm the prior

- The goal for an uninformative prior is to minimally impact the posterior relative to how much the data is impacting the prior.

12

Conjugate Priors for Normal Likelihood

Let's go through an example, the normal likelihood with unknown mean and variance.

$$y|\mu,\sigma^2 \sim N(\mu,\sigma^2)$$

This is our likelihood. WRITE THIS FOR THE MIDTERM

We are going to use a slightly different construction than what you are used to seeing for conjugate priors.

this is a heirarchical prior

From a frequentist standpoint, this looks like a sample distr.

$$\begin{cases} \mu|\sigma^2 \sim N(\mu_0, \dfrac{\sigma^2}{\kappa_0}) \\ \sigma^2 \sim Inv\chi^2(\nu_0, \sigma_0^2) \end{cases}$$

First is our conditional parameter. We choose here Mu0 and Kappa0

This is an inverse Chi, square. It has v0 (Nu0) and sigma squared 0. We have 4 hyper parameters

Hyperparameters:

- μ_0- Prior mean
- σ_0^2 - Prior variance
- κ_0 - Prior "sample size"
- ν_0 - Prior degrees of freedom

13

Conjugate Priors for Normal Likelihood

Posterior parameter values below

Q: figure out what is an informative and uninformative prior selection for the example.

-Note: you should choose hyperparameter values that minimally affect the stuff on the left side. this is only for uninformative priors.
-anything with the subscript 0 are our hyperparameters.

Write this for the exam

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2$$

How would you chose your prior hyper parameters to be minimally informative?

- Choose values that maximize how much the data is able to influence the posterior.

14

Samplers

When we don't have conjugacy, we need to approximate the posterior using samplers (or variational inference, but let's go with samplers here.)

General idea of all samplers:

- Take a guess at what the posterior parameter value should be (ala, sample it).
- Accept or reject that guess based on how well it fits the posterior density.
- Fix that parameter value, and move to the next parameter. Repeat until convergence.

Three main types of samplers: These three will be on the exam

1. Gibbs Sampler
2. Metropolis/Metropolis-Hastings
3. Hamiltonian MCMC/NUTS

15

Gibbs Sampler

Gibbs samplers are used when the *conditional posteriors* have nice distributions (technically not the same thing as conjugacy, but in practice, I like to call this conjugacy.)

For example, if $\mu|\sigma^2 \sim Normal$, while $\sigma^2|\mu$ is distributed as, say, a Gamma, but the joint distribution μ, σ^2 has an unknown distribution, then we can Gibbs sample.

Gibbs sampling is simple:

1. Sample $\mu|\sigma^2 \rightarrow \mu^*$
2. Sample $\sigma^2|\mu^* \rightarrow \sigma^{2*}$
3. Repeat

the goal is to get the full posterior.

The advantages of Gibbs are that a) its computationally fast (no need to reject guesses). The disadvantage of Gibbs is that it requires the models/priors to have this conditional conjugacy property.

you have to choose your models and priors correctly for Gibbs

16

Metropolis-Hastings

Will be on the exam. formulas not required, but you can use theory to explain.
Just checking conceptual understanding of sampler.
advantage: approach that always works
disadvantage: doesn't work in our lifetime. Other than that, nice theoretical guarantess

Metropolis-Hastings works for any type of model/prior choice (exceptions do apply).

Conceptually, Metropolis-Hastings works like this:

1. Throw out a guess for your parameter value using a *proposal distribution.*
2. Evaluate if that guess is better or worse than the value you were standing on.
    1. If better, then move to your guess.
    2. If worse, move to your guess with inverse probability of how much worse your guess is.

we are trying to get to the top of the hill, but once there, we want to walk around. In other words, you want to move closer if your guess is better, and stay where you are if worse

3. Repeat until you have converged to the posterior.

The Hastings component of Metropolis-Hastings is an adjustment to step 2 that accounts for oddly shaped proposal distributions.

The advantage of MH is that it, in theory, will always work.

The disadvantage is that it only works in practice if you have tuned it correctly.

- It might take too long to sample, even if it is guaranteed to eventually converge.

17

Hamiltonian/NUTS

similar to metropolis-hastings

Hamiltonian MCMC/NUTS is similar to Metropolis-Hastings, but uses information about the posterior "landscape" to help guide guesses.

- When it takes a guess at the parameter values, it evaluates the gradient (slope) of the posterior landscape, and nudges the guess towards a higher probability value.
- Metaphor: You are trying to find the top of a hill in the dark. You throw out a sensor, and it tells you the new location is higher. You go to that location and feel the slope of the hill. You then walk a couple more steps up hill.

NUTS comes in and backs us up a little bit.

Advantages:

- Very effective sampling of high dimensional parameter sets.
- NUTS sampler eliminates the need for tuning.

Disadvantages:

- Requires continuous parameters to get gradient information

18

Midterm on Tuesday

Pen and Paper Midterm 10/22

- One (regular sized) sheet of notes, back and front.
- Calculator is okay, you technically shouldn't need it.
  - If you don't have a calculator, you can use your phone calculator. Don't cheat!

Remember, the midterm is only worth 10% of your grade. You could skip the midterm and still get 100% in the class (doing extra credit on the assignments).

- Review these slides, they cover what is going to be on the midterm.
- Review lecture slides generally, refresh your understanding of conjugate priors and what it means to be uninformative/informative.
- Review laws of probability (use ChatGPT to come up with practice problems!)

19