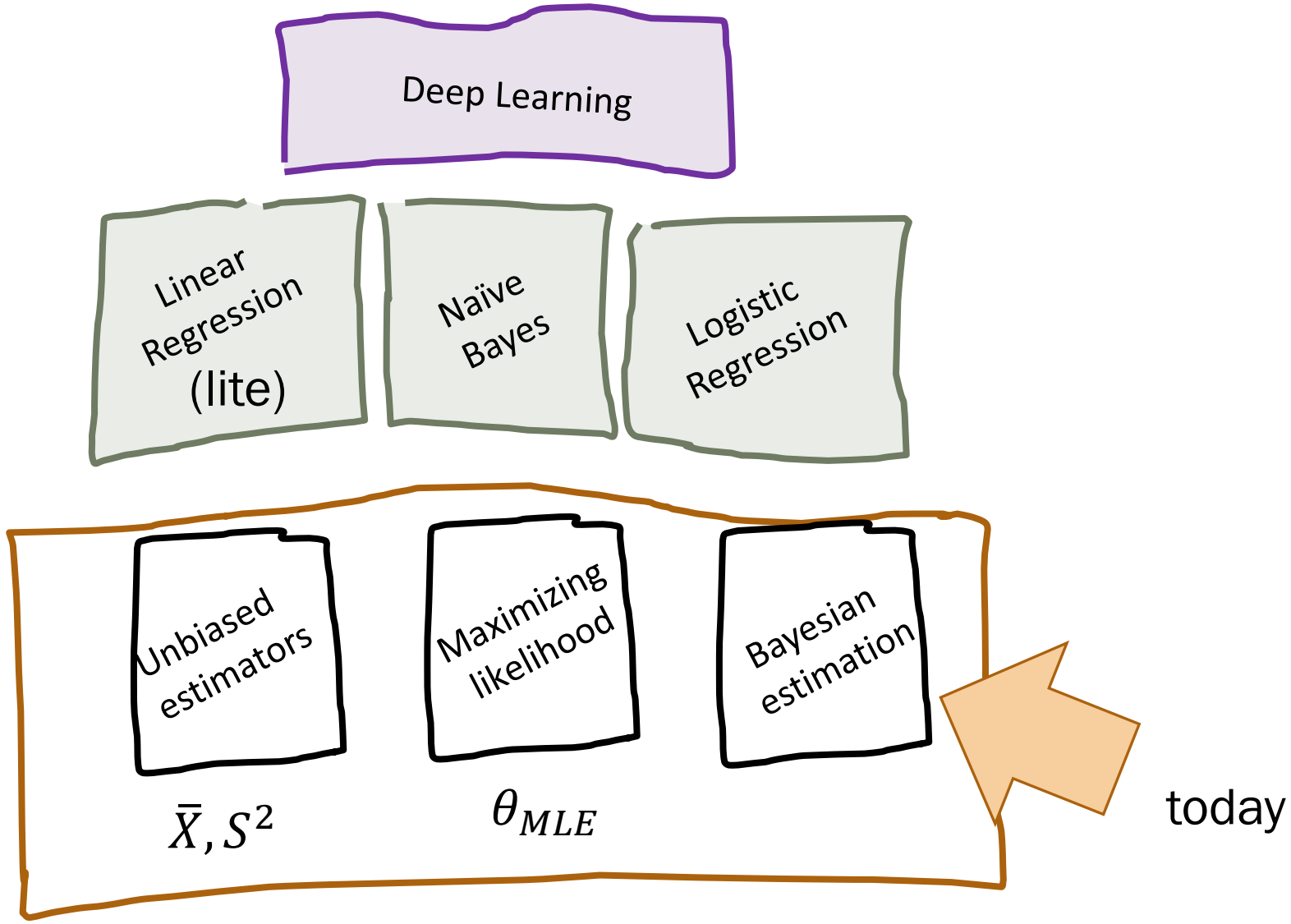
23: Maximum A Posteriori (MAP)

David Varodayan
March 2, 2020
Adapted from slides by Lisa Yan

Our path


Deep Learning
Linear Regression (lite)
Naïve Bayes
Logistic Regression
Unbiased estimators
Maximizing likelihood
Bayesian estimation
$\bar{X}, S^2$
θ_{MLE}
today

Today's plan

Maximum Likelihood

- MLE for Multinomial

Maximum A Posteriori

- MAP for Binomial with Beta Conjugate Prior
- MAP for Multinomial with Beta Conjugate Prior
- MAP for Poisson with Gamma Conjugate Prior

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \ldots, Y_n$.

- Let $Y_k \sim \text{Multinomial}(p_1, p_2, \ldots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Example: Suppose Y_k = outcome of 6-sided die.

$m = 6, \sum_{i=1}^{6} p_i = 1$

- Roll the dice $n = 12$ times.
- Observe data: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

$$X_1 = 3, X_2 = 2, X_3 = 0,$$
$$X_4 = 3, X_5 = 1, X_6 = 3$$

Check: $X_1 + X_2 + \cdots + X_6 = 12$

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \ldots, Y_n$.

- Let $Y_k \sim$ Multinomial$(p_1, p_2, \ldots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Joint PDF $f(X_1, X_2, \ldots, X_m | p_1, p_2, \ldots, p_m)$:

A. $\frac{n!}{x_1!\, x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$

B. $p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$

C. $\frac{n!}{x_1!\, x_2! \cdots x_m!} x_1^{p_1} x_2^{p_2} \cdots x_m^{p_m}$

Likelihood $L(\theta)$
of observing the sample (size n)
$(X_1, X_2, \ldots, X_m)$

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \ldots, Y_n$.

- Let $Y_k \sim \text{Multinomial}(p_1, p_2, \ldots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Joint PDF $f(X_1, X_2, \ldots, X_m | p_1, p_2, \ldots, p_m) = \frac{n!}{x_1!\, x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} = L(\theta)$

Log-likelihood:

$$LL(\theta) = \log(n!) - \sum_{i=1}^{m} \log(X_i!) + \sum_{i}^{m} X_i \log(p_i), \text{ such that } \sum_{i=1}^{m} p_i = 1$$

Optimize with Lagrange multipliers in extra slides

θ_{MLE}: $p_i = \frac{X_i}{n}$

Intuitively, probability p_i = proportion of outcomes

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

What is θ_{MLE}? (select all that apply)

A. $p_1 = 3/12$
B. $p_2 = 2/12$
C. $p_3 = 0/12$
D. $p_4 = 3/12$
E. $p_5 = 1/12$
F. $p_6 = 3/12$
G. Other

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

θ_{MLE}:

$$p_1 = 3/12$$
$$p_2 = 2/12$$
$$p_3 = 0/12$$
$$p_4 = 3/12$$
$$p_5 = 1/12$$
$$p_6 = 3/12$$

- MLE: you'll **never...EVER...** roll a three.
- Do you really believe that?

Frequentist:	Bayesian:
Roll more! Prob. = frequency in limit	Have prior beliefs of probability, even before any rolls!

Announcements

Problem Set 6

Due:	Wednesday 3/11
Covers:	Up to Lecture 25
Extra Python Office Hours:	Saturday 3/7, 3-5PM

Autograded Coding Problems

Run your code in the command line, not just in a Jupyter notebook cell

Late Day Reminder

No late days permitted past last day of the quarter, 3/13

Today's plan

Maximum Likelihood
- MLE for Multinomial

Maximum A Posteriori
- MAP for Binomial with Beta Conjugate Prior
- MAP for Multinomial with Beta Conjugate Prior
- MAP for Poisson with Gamma Conjugate Prior

Estimating our parameter directly

(our focus so far)

Maximum Likelihood Estimator (MLE)

What is the parameter θ that **maximizes the likelihood** of our observed data $(x_1, x_2, \dots, x_n)$?

$$L(\theta) = f(X_1, X_2, \dots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} f(X_1, X_2, \dots, X_n|\theta)$$

likelihood of data

(our focus today)

Maximum a Posteriori (MAP) Estimator

Given our observed data $(x_1, x_2, \dots, x_n)$, what is the **most likely parameter** θ?

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \dots, X_n)$$

posterior distribution of θ

Maximum A Posterior (MAP) Estimator

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$ (data).

def The **Maximum a Posteriori (MAP) Estimator** of θ is the value of θ that maximizes the posterior distribution of θ.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

Intuition with Bayes' Theorem:

$L(\theta)$, probability of data given parameter θ

After seeing data, posterior belief of θ

$$\underbrace{P(\theta|\text{data})}_{\text{posterior}} = \frac{\overbrace{P(\text{data}|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(\text{data})}$$

Before seeing data, prior belief of θ

Solving for θ_{MAP}

$$\text{posterior } P(\theta|\text{data}) = \frac{\text{likelihood } P(\text{data}|\theta) \quad \text{prior } P(\theta)}{P(\text{data})}$$

- Observe data: $X_1, X_2, \ldots, X_n$, all i.i.d.
- Let likelihood be same as MLE: $f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\,\theta)$
- Let the prior distribution of θ be $g(\theta)$.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n) = \arg\max_{\theta} \frac{f(X_1, X_2, \ldots, X_n|\theta)g(\theta)}{h(X_1, X_2, \ldots, X_n)} \quad \text{(Bayes' Theorem)}$$

$$= \arg\max_{\theta} \frac{g(\theta)\prod_{i=1}^{n} f(X_i|\,\theta)}{h(X_1, X_2, \ldots, X_n)} \quad \text{(independence)}$$

$$= \arg\max_{\theta} g(\theta)\prod_{i=1}^{n} f(X_i|\,\theta) \quad (1/h(X_1, X_2, \ldots, X_n) \text{ is a positive constant w.r.t. } \theta)$$

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\,\theta)\right)$$

Maximum A Posterior (MAP) Estimator

The MAP estimator has 2 interpretations:

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i | \theta) \right)$$

The mode of the posterior distribution of θ

The θ that maximizes log prior + log-likelihood

In both cases, you must specify your prior, $g(\theta)$.

Key to MAP estimator: You should pick a prior $g(\theta)$ that makes computing the mode of the posterior distribution is **easy**.

(in this class) Use a conjugate distribution.

Today's plan

Maximum Likelihood

- MLE for Multinomial

Maximum A Posteriori

- MAP for Binomial with Beta Conjugate Prior
- MAP for Multinomial with Beta Conjugate Prior
- MAP for Poisson with Gamma Conjugate Prior

Beta distribution refresher

We have seen one conjugate distribution so far:

$$X \sim \text{Beta}(a, b)$$

$$a > 0, b > 0$$

Support of X: $(0, 1)$

PDF $$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$, normalizing constant

- Beta is the **conjugate distribution** for Bernoulli, meaning:

Prior	$\text{Beta}(a = n_{imag} + 1, b = m_{imag} + 1)$
Experiment	Observe n successes and m failures
Posterior	$\text{Beta}(a = n_{imag} + n + 1, b = m_{imag} + m + 1)$

- Mode of $\text{Beta}(a, b)$: $\frac{a-1}{a+b-2}$

MAP estimator for Binomial

Suppose you observe data D: n heads, m tails

1. Decide model. Binomial with parameter p
2. Decide prior distribution of parameter θ, $g(\theta)$. $\theta \sim \text{Beta}(a+1, b+1)$
3. Compute θ_{MAP} (below)

Solution:

- Beta is a conjugate distribution for Binomial.
- If prior $\theta \sim \text{Beta}(a+1, b+1)$ and data = $\{n$ heads, m tails$\}$, then posterior distribution
$$\theta|\text{data} \sim \text{Beta}(a+n+1, b+m+1)$$
- θ_{MAP} is mode of posterior distribution $\theta_{MAP} = \dfrac{a+n}{a+n+b+m}$

(mode of $\text{Beta}(a+n+1, b+m+1)$)

MAP estimator for Binomial, from first principles

Suppose you observe data D: n heads, m tails

1. Decide model. Binomial with parameter p
2. Decide prior distribution of parameter θ, $g(\theta)$. $\theta \sim \text{Beta}(a+1, b+1)$
3. Compute θ_{MAP} (below)

$$\theta_{MAP} = \arg\max_{\theta}(\log g(\theta) + \log f(X_1, X_2, \ldots, X_n|\theta))$$

(θ_{MAP} = argmax of log prior + log-likelihood)

$$= \arg\max_{p}\left(\log\left(\frac{1}{\beta}p^{a+1-1}(1-p)^{b+1-1}\right) + \log\left(\binom{n+m}{n}p^n(1-p)^m\right)\right)$$

(PDF of Beta, likelihood of n heads, m tails)

$$= \arg\max_{p}\left(\log\frac{1}{\beta} + a\log(p) + b\log(1-p) + \log\binom{n+m}{n} + n\log p + m\log(1-p)\right)$$

$$= \arg\max_{p}\big((a+n)\log(p) + (b+m)\log(1-p)\big)$$

(eliminate constants w.r.t. $\arg\max_p$)

MAP estimator for Binomial, from first principles

Suppose you observe data D: n heads, m tails

1. Decide model. Binomial with parameter p
2. Decide prior distribution of parameter θ, $g(\theta)$. $\theta \sim \text{Beta}(a+1, b+1)$
3. Compute θ_{MAP} (below)

$$\theta_{MAP} = p_{MAP} = \arg\max_p \left((a+n)\log(p) + (b+m)\log(1-p) \right)$$

Differentiate w.r.t. p and set to 0:

$$\frac{(a+n)}{p} - \frac{(b+m)}{1-p} = 0$$

Solve for p:

$$(a+n)(1-p) = (b+m)p$$
$$(a+n) - (a+n)p = (b+m)p$$
$$p(a+n+b+m) = a+n$$

$$p_{MAP} = \frac{a+n}{a+n+b+m}$$

How does MAP work?

0. Observe data — n heads, m tails
1. Choose model — Binomial(p)
2. Choose prior of θ
   - (some $g(\theta)$)
   - (choose conjugate) Beta($a+1, b+1$)
3. Compute posterior of θ given data
   - $\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\theta)$
   - Beta($a+n+1, b+m+1$)
4. $\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n)$
   - Differentiate
   - Solve

This is nice! Mode of Beta: $\frac{a+n}{a+n+b+m}$

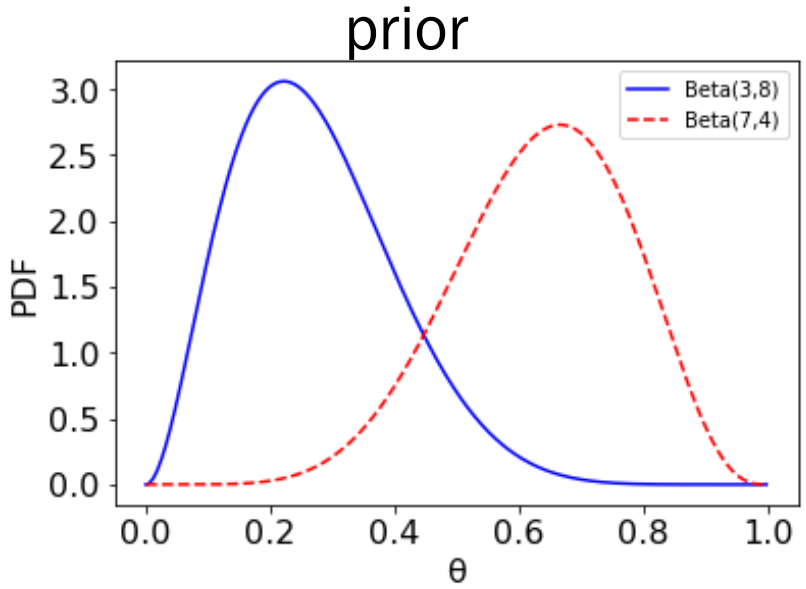
How does MAP work?

0. Observe data — n heads, m tails
1. Choose model — Binomial(p)
2. Choose prior of θ — (some $g(\theta)$) / (choose conjugate) Beta($a+1, b+1$)
3. Compute posterior of θ given data — $\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\theta)$ / Beta($a+n+1, b+m+1$)
4. $\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n)$
   - Differentiate
   - Solve

Mode of Beta: $\frac{a+n}{a+n+b+m}$

Where'd you get them priors?

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
  - Prior 1: Beta(3,8): 2 imaginary heads, 7 imaginary tails — mode: $\frac{2}{9}$
  - Prior 2: Beta(7,4): 6 imaginary heads, 3 imaginary tails — mode: $\frac{6}{9}$



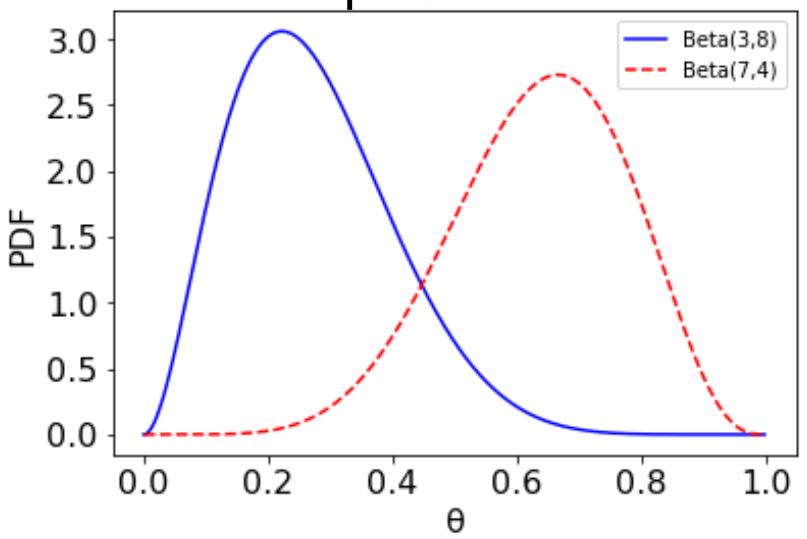
Now flip 100 coins and get 58 heads and 42 tails.

1. What are the two posterior distributions?
2. What are the modes of the two posterior distributions?

Where'd you get them priors?

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
  - Prior 1: Beta(3,8): 2 imaginary heads, 7 imaginary tails — mode: $\frac{2}{9}$
  - Prior 2: Beta(7,4): 6 imaginary heads, 3 imaginary tails — mode: $\frac{6}{9}$

prior



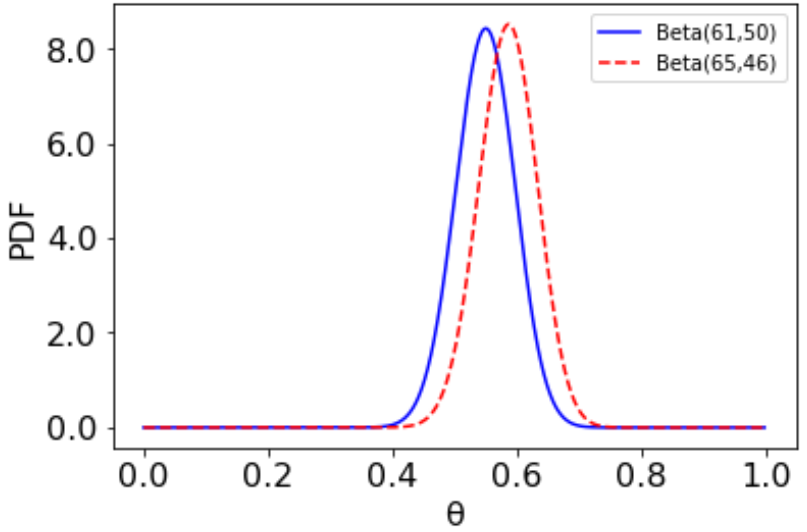
Now flip 100 coins and get 58 heads and 42 tails.

Posterior 1: Beta(61,50) — mode: $\frac{60}{109}$

Posterior 2: Beta(65,46) — mode: $\frac{64}{109}$

As long as we collect enough data, posteriors will converge to the true value.

posterior



MAP estimator so far

MAP estimator: $\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n)$

The mode of the posterior distribution of θ

You should pick a prior $g(\theta)$ that makes computing the mode of the posterior distribution **easy**.

Use a conjugate distribution.

The conjugate for Binomial is Beta.

- Our **prior** (subjective) belief:
  $\theta \sim \text{Beta}(a+1, b+1)$
  (Saw $a+b=$ imaginary trials; of those, a were successes.)
- **Posterior** distribution:
  $(\theta|n \text{ heads}, m \text{ tails}) \sim$
  $\text{Beta}(a+n+1, b+m+1)$

Conjugate distributions

MAP estimator:

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n)$$

The mode of the posterior distribution of θ

Distribution parameter	Prior distribution for parameter
Bernoulli p	Beta
Binomial p	Beta
Multinomial p_i	Dirichlet
Poisson λ	Gamma
Exponential λ	Gamma
Normal μ	Normal
Normal σ^2	Inverse Gamma

Today's plan

Maximum Likelihood

- MLE for Multinomial

Maximum A Posteriori

- MAP for Binomial with Beta Conjugate Prior
- MAP for Multinomial with Beta Conjugate Prior
- MAP for Poisson with Gamma Conjugate Prior

MAP for Multinomial

Dirichlet$(a_1, a_2, \ldots, a_m)$ is the conjugate for Multinomial.

- Generalizes Beta in the same way Multinomial generalizes Bernoulli/Binomial:

$$f(x_1, x_2, \ldots, x_m) = \frac{1}{B(a_1, a_2, \ldots, a_m)} \prod_{i=1}^{m} x_i^{a_i - 1}$$

Prior Dirichlet$(a_1 + 1, a_2 + 1, \ldots, a_m + 1)$
Saw $\sum_{i=1}^{m} a_i$ imaginary trials, a_i of outcome i

Experiment Observe $n_1 + n_2 + \cdots + n_m$ new trials, with n_i of outcome i

Posterior Dirichlet$(a_1 + n_1 + 1, a_2 + n_2 + 1, \ldots, a_m + n_m + 1)$

MAP:
$$p_i = \frac{a_i + n_i}{\sum_{i=1}^{m} a_i + \sum_{i=1}^{m} n_i}$$

Laplace smoothing

MAP with **Laplace smoothing**: a prior which represents **one** imagined observation of each outcome.

Consider our previous 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

Recall θ_{MLE}:

$p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$
$p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

θ_{MAP} with Laplace smoothing:

- Assume Dirichlet prior where each outcome seen $k = 1$ times.
- **Laplace estimate**:

$$p_i = \frac{X_i + 1}{n + m}$$

$p_1 = 4/18, p_2 = 3/18, p_3 = 1/18,$
$p_4 = 4/18, p_5 = 2/18, p_6 = 4/18$

Laplace smoothing avoids the case where you estimate a parameter of 0.

Today's plan

Maximum Likelihood

- MLE for Multinomial

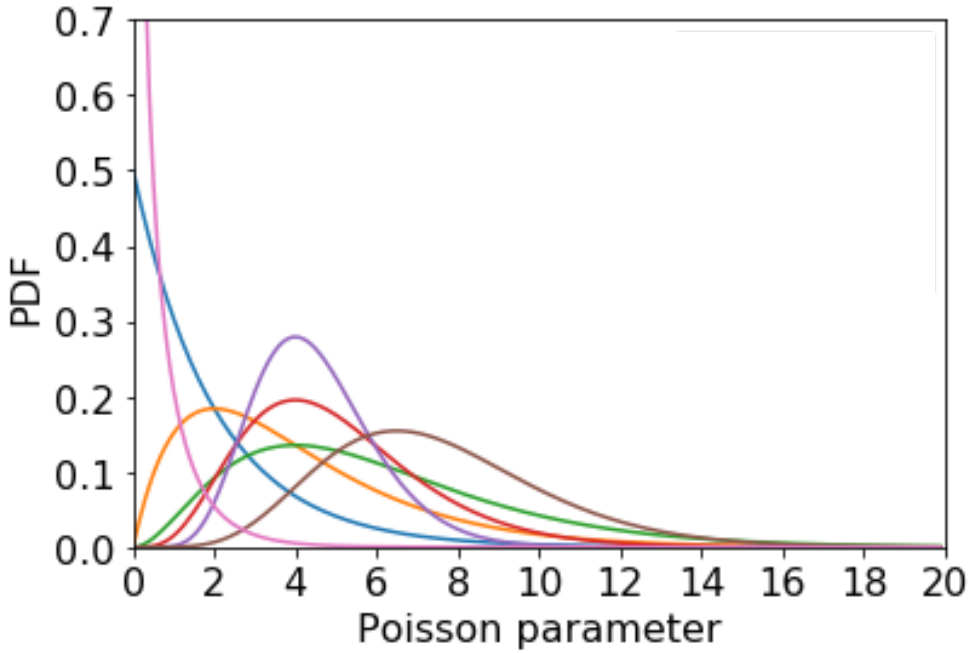
Maximum A Posteriori

- MAP for Binomial with Beta Conjugate Prior
- MAP for Multinomial with Beta Conjugate Prior
- MAP for Poisson with Gamma Conjugate Prior

Gamma Distribution

Gamma(α, β) is the conjugate for Poisson.

- Also conjugate for Exponential, but we won't delve into that
- Mode of gamma: α/β



Prior $\lambda \sim$ Gamma(α, β)
Saw α total imaginary events during β prior time periods

Experiment Observe n events during next k time periods

Posterior $(\lambda | n$ events in k periods$)$
$\sim$Gamma$(\alpha + n, \beta + k)$

$$\theta_{MAP} = \frac{\alpha + n}{\beta + k}$$

MAP for Poisson

Gamma(α, β) is conjugate for Poisson

Mode: α/β

1. What does it mean to have a prior of $\lambda \sim \text{Gamma}(10,5)$?

MAP for Poisson

Gamma(α, β) is conjugate for Poisson

Mode: α/β

1. What does it mean to have a prior of $\lambda \sim \text{Gamma}(10,5)$?

Observe 10 imaginary events in 5 time periods, i.e., observe at Poisson rate $\lambda = 2$

Now perform the experiment and see 11 events in next 2 time periods.

2. Given your prior, what is the posterior distribution?

3. What is λ_{MAP}?

Extra slides

Finding the MLE for Multinomial

MLE for Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \ldots, Y_n$.

- Let $Y_k \sim$ Multinomial$(p_1, p_2, \ldots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Joint PDF $f(X_1, X_2, \ldots, X_m | p_1, p_2, \ldots, p_m) = \frac{n!}{x_1!\, x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} = L(\theta)$

Log-likelihood:

$$LL(\theta) = \log(n!) - \sum_{i=1}^{m} \log(X_i!) + \sum_{i}^{m} X_i \log(p_i), \text{ such that } \sum_{i=1}^{m} p_i = 1$$

Optimize with Lagrange multipliers in extra slides

θ_{MLE}: $p_i = \frac{X_i}{n}$

Intuitively, probability p_i = proportion of outcomes

Optimizing MLE for Multinomial

$\theta = (p_1, p_2, \ldots, p_m)$

$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$, where $\sum_{i=1}^{m} p_i = 1$

Use Lagrange multipliers to account for constraint

Lagrange multipliers: $A(\theta) = LL(\theta) + \lambda\left(\sum_{i=1}^{m} p_i - 1\right) = \sum_{i}^{m} X_i \log(p_i) + \lambda\left(\sum_{i=1}^{m} p_i - 1\right)$ (drop non-p_i terms)

Differentiate w.r.t. each p_i, in turn: $\frac{\partial A(\theta)}{\partial p_i} = X_i \frac{1}{p_i} + \lambda = 0 \Rightarrow p_i = -\frac{X_i}{\lambda}$

Solve for λ, noting $\sum_{i=1}^{m} X_i = n, \sum_{i=1}^{m} p_i = 1$: $\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} -\frac{X_i}{\lambda} = 1 \Rightarrow 1 = -\frac{n}{\lambda} \Rightarrow \lambda = -n$

Substitute λ into p_i $p_i = \frac{X_i}{n}$