Bayesian Methods

Rishabh Iyer
University of Texas at Dallas

based on the slides of Vibhav Gogate and Nick Rouzzi

Binary Variables

$$E[X] = \sum_{x \in D} p(X=x) \cdot x$$

- Coin flipping: heads=1, tails=0 with bias μ

$$p(X = 1|\mu) = \mu$$

$$Var(X) = E[X-\mu]^2$$
$$= \sum_{x \in D} [x-\mu]^2 p(X=x)$$

- Bernoulli Distribution

$$Bern(x|\mu) = \mu^x \cdot (1-\mu)^{1-x}$$

$$E[X] = \mu = 0 \cdot (1-\mu) + 1 \cdot \mu$$

$$var(X) = \mu \cdot (1-\mu)$$

$$Bern(1|\mu) = p(X=1|\mu) = \mu$$
$$Bern(0|\mu) = p(X=0|\mu) = 1-\mu$$

$$\mathrm{Var}(X) = E[(X-\mu)^2], \quad \mu = E[X]$$

$$= E[X^2 - 2X\mu + \mu^2]$$

$$= E[X^2] - 2(E[X])^2 + (E[X])^2$$

$$= E[X^2] - (E[X])^2$$

For Bern:

$$E[X^2] = 0^2(1-\mu) + 1^2 \cdot \mu = \mu$$

$$(E[X])^2 = \mu^2$$

$$\mathrm{Var}(X) = \mu - \mu^2$$

Binary Variables

- N coin flips: $X_1, \dots, X_N$

$$p(\sum_i X_i = m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- Binomial Distribution

$$Bin(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

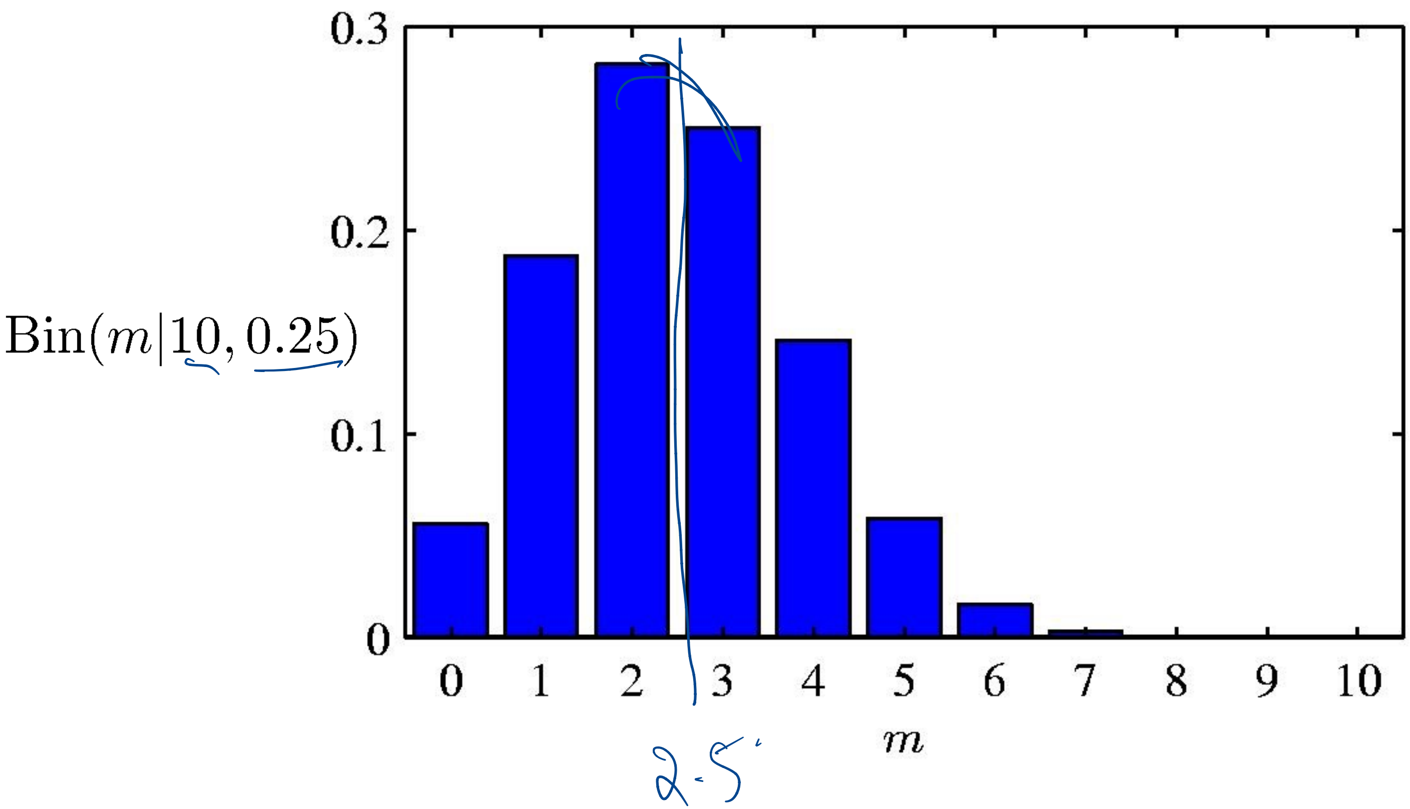
heads = m

$$E\left[\sum_i X_i\right] = N\mu$$

$$var\left[\sum_i X_i\right] = N\mu(1-\mu)$$

Binomial Distribution

$N = 10, \mu = 0.25$



Estimating the Bias of a Coin

- Suppose that we have a coin, and we would like to figure out what the probability is that it will flip up heads
  - How should we estimate the bias?

Estimating the Bias of a Coin

- Suppose that we have a coin, and we would like to figure out what the probability is that it will flip up heads
  - How should we estimate the bias?

T H T H H

  - With these coin flips, our estimate of the bias is: ?

$$\mu = \frac{3}{5}$$

Estimating the Bias of a Coin

- Suppose that we have a coin, and we would like to figure out what the probability is that it will flip up heads
  - How should we estimate the bias?

  - With these coin flips, our estimate of the bias is: 3/5
    - Why is this a good estimate?

Coin Flipping – Binomial Distribution

D

- $P(Heads) = \theta, \; P(Tails) = 1 - \theta$
- Flips are i.i.d.
  - Independent events
  - Identically distributed according to Binomial distribution
- Our training data consists of α_H heads and α_T tails

$$p(D|\theta) = \theta^{\alpha_H} \cdot (1 - \theta)^{\alpha_T}$$

Maximum Likelihood Estimation (MLE)

- **Data:** Observed set of α_H heads and α_T tails
- **Hypothesis:** Coin flips follow a Bernoulli distribution
- **Learning:** Find the "best" θ
- **MLE:** Choose θ to maximize probability of D given θ

$$
\begin{aligned}
\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\
&= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta)
\end{aligned}
$$

First Parameter Learning Algorithm

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

Set derivative to zero, and solve!

$$\begin{aligned}\frac{d}{d\theta}\ln P(\mathcal{D} \mid \theta) &= \frac{d}{d\theta}[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}] \\ &= \frac{d}{d\theta}[\alpha_H \ln \theta + \alpha_T \ln(1-\theta)] \\ &= \alpha_H \frac{d}{d\theta}\ln \theta + \alpha_T \frac{d}{d\theta}\ln(1-\theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} \quad = 0\end{aligned}$$

First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

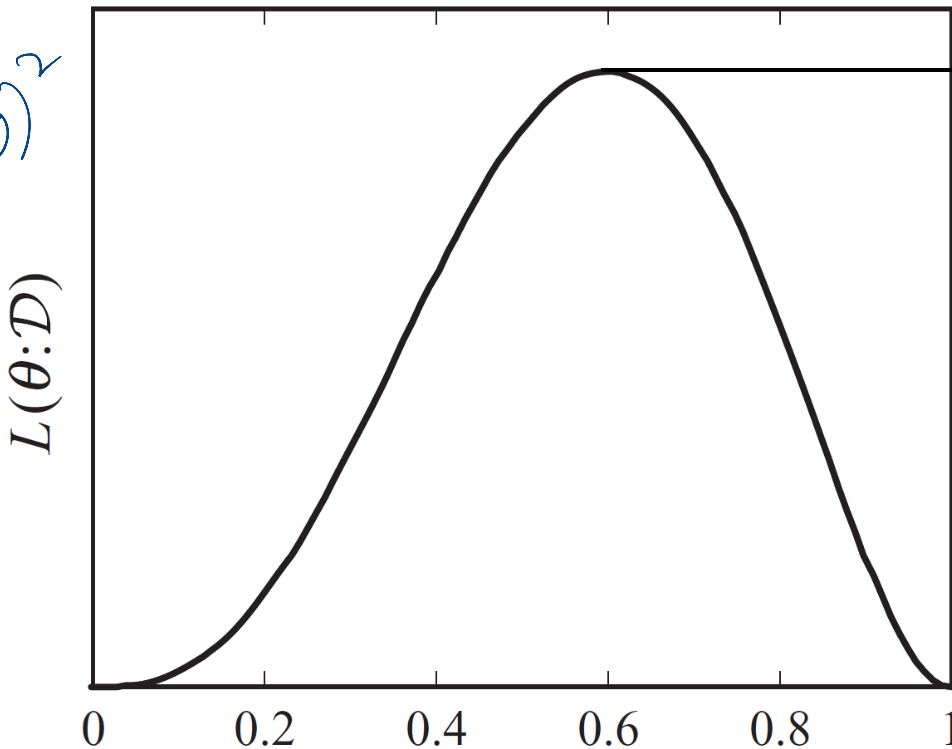
Set derivative to zero, and solve!

$$\begin{aligned}\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) &= \frac{d}{d\theta}\left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right] \\ &= \frac{d}{d\theta}\left[\alpha_H \ln \theta + \alpha_T \ln(1-\theta)\right] \\ &= \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1-\theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} \quad = 0\end{aligned}$$

$$\boxed{\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

Coin Flip MLE

$\theta^2(1-\theta)^2$



$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Priors

- Suppose we have 5 coin flips all of which are heads
    - Our estimate of the bias is?

Priors

- Suppose we have 5 coin flips all of which are heads
  - MLE would give $\theta_{MLE} = 1$
  - This event occurs with probability $\frac{1}{2^5} = \frac{1}{32}$ for a fair coin
  - Are we willing to commit to such a strong conclusion with such little evidence?

Priors

- Priors are a Bayesian mechanism that allow us to take into account “prior” knowledge about our belief in the outcome

- Rather than estimating a single θ, consider a distribution over possible values of θ given the data

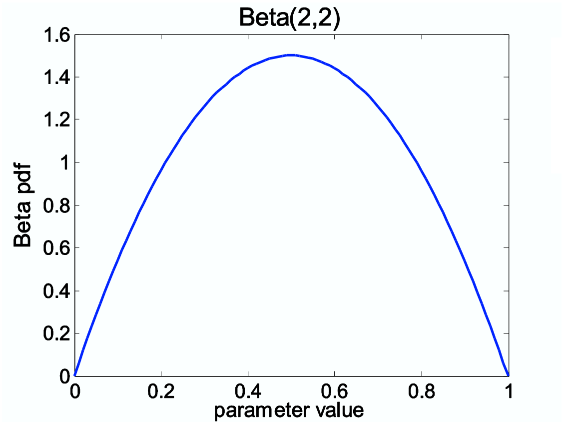
  $p(D|\theta) \leftarrow$ MLE

  $p(\theta|D)$

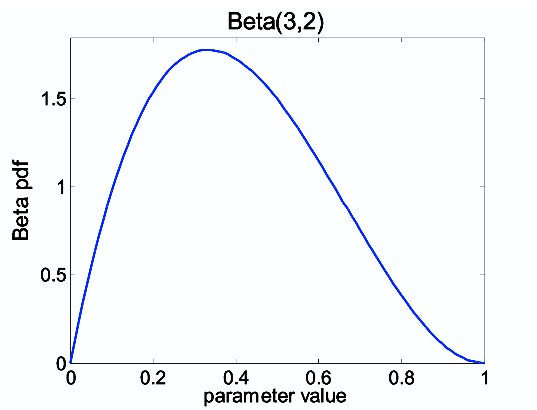
  - Update our prior after seeing data

Our best guess in the absence of any data

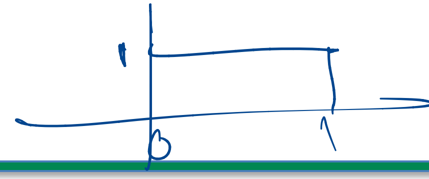
Our estimate after we see some data



Observe flips e.g.: {tails, tails}

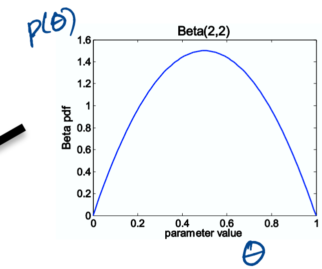


Bayesian Learning



Apply Bayes rule:

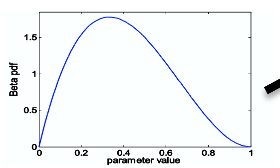
Prior



Data Likelihood

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Posterior



Normalization

- Or equivalently: $p(\theta|D) \propto p(D|\theta)p(\theta)$
- For uniform priors this reduces to the MLE objective

$$p(\theta) \propto 1 \qquad \Rightarrow \qquad p(\theta|D) \propto p(D|\theta)$$

Picking Priors

- How do we pick a good prior distribution?
  - Could represent expert domain knowledge
  - Statisticians choose them to make the posterior distribution “nice” (conjugate priors)
- What is a good prior for the bias in the coin flipping problem?

Picking Priors

- How do we pick a good prior distribution?
  - Could represent expert domain knowledge
  - Statisticians choose them to make the posterior distribution “nice” (conjugate priors)
- What is a good prior for the bias in the coin flipping problem?
  - Truncated Gaussian (tough to work with)
  - Beta distribution (works well for binary random variables)

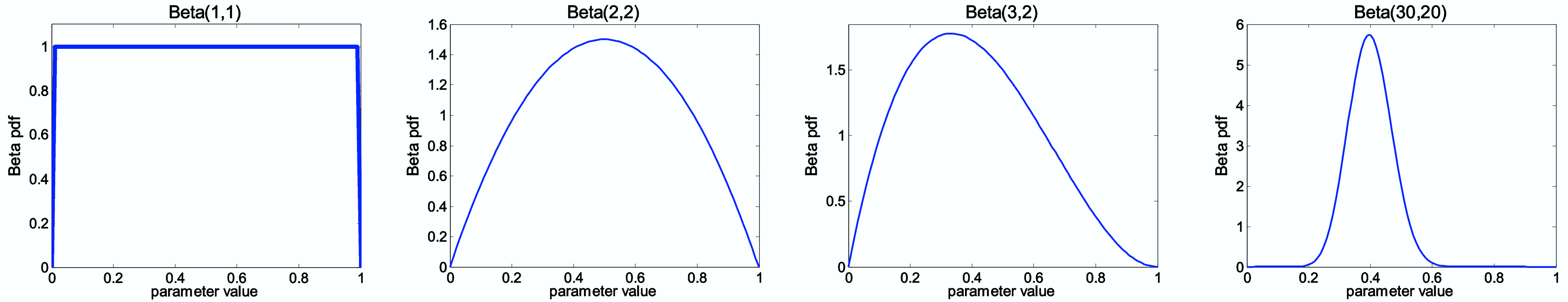
Coin Flips with Beta Distribution

Likelihood function:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Prior:

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



$$\begin{aligned}P(\theta \mid \mathcal{D}) &\propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\ \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1} \\ &= \theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1} \\ &= Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)\end{aligned}$$

MAP Estimation

- Choosing θ to maximize the posterior distribution is called maximum a posteriori (MAP) estimation

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D)$$

- The only difference between θ_{MLE} and θ_{MAP} is that one assumes a uniform prior (MLE) and the other allows an arbitrary prior

Priors

- Suppose we have 5 coin flips all of which are heads
  - MLE would give $\theta_{MLE} = 1$
  - MLE with a $Beta(2,2)$ prior gives $\theta_{MAP} = \frac{6}{7} \approx .857$
  - As we see more data, the effect of the prior diminishes
    - $\theta_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \approx \frac{\alpha_H}{\alpha_H + \alpha_T}$ for large # of observations

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound

    - Suppose $Y_1, \ldots, Y_N$ are i.i.d. random variables taking values in $\{0, 1\}$ such that $E_p[Y_i] = y$. For $\epsilon > 0$,

$$p\left(\left|y - \frac{1}{N}\sum_i Y_i\right| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound

  - For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

$$p\left(\left|\theta_{true} - \frac{1}{N}\sum_i X_i\right| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound
  - For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

$$p(|\theta_{true} - \theta_{MLE}| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound

  - For the coin flipping problem with $X_1, \dots, X_n$ iid coin flips and $\epsilon > 0$,

$$p(|\theta_{true} - \theta_{MLE}| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

$$\delta \geq 2e^{-2N\epsilon^2} \Rightarrow N \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$$