

# CS 4824/ECE 4424: MLE & MAP

## *Acknowledgement:*

Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

# Principles for estimating probabilities

$\Theta$  = set of model parameters

$$\Theta = \hat{P}(X=1)$$

$$\hat{P}(X=0) = (1 - \theta)$$

$\alpha_1 = \# \text{ heads}$   
 $\alpha_0 = \# \text{ tails}$

- **Principle 1.** Maximum likelihood estimation (MLE)

- Choose parameter  $\theta$  that maximizes  $P(\text{data} | \theta)$

- $$\hat{\theta} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

$$\text{MLE} = \arg \max_{\theta} P(D | \theta)$$

- **Principle 2.** Maximum a posteriori estimation (MAP)

- Choose parameter  $\theta$  that maximizes  $P(\theta | \text{data})$

- $$\hat{\theta} = \frac{(\alpha_1 + \beta_1)}{(\alpha_1 + \beta_1) + (\alpha_0 + \beta_0)}$$

$$\text{MAP} = \arg \max_{\theta} P(\theta | \text{data})$$

$\beta_1 = \# \text{ hallucinated heads}$

$\beta_0 = \# \text{ hallucinated tails}$

# Formal treatment

D

- **Principle 1.** Maximum likelihood estimation (MLE)

- $P(X = 1) = \theta; P(X = 0) = 1 - \theta$

*independent and identically distributed (iid)*

- Data D : 1 0 0 1 1

- $P(D | \theta) : \theta (1 - \theta) (1 - \theta) \theta \theta = \theta^{\alpha_1} (1 - \theta)^{\alpha_0}$

$\theta^3 (1-\theta)^2$

Flip produces data D with  $\alpha_1$  heads and  $\alpha_0$  tails, iid ~ Bernoulli

$\alpha_1$  heads and  $\alpha_0$  are counts that sum these outcomes ~ Binomial

Learning  $\theta$  is an optimization problem. What's the objective function?

- $\hat{\theta} = \arg \max_{\theta} P(D | \theta) = \arg \max_{\theta} \ln P(D | \theta)$

# Derivation

$$P(D | \theta) = \theta^{\alpha_1} (1 - \theta)^{\alpha_0}$$

$$\hat{\theta} = \arg \max_{\theta} P(D | \theta) = \arg \max_{\theta} \ln P(D | \theta)$$

$$\hat{\theta} = \arg \max_{\theta} \ln [\theta^{\alpha_1} (1 - \theta)^{\alpha_0}]$$

$$= \arg \max_{\theta} \alpha_1 \ln \theta + \alpha_0 \ln (1 - \theta)$$

→ Set the derivative to zero  $\frac{d}{d\theta} \ln P(D | \theta) = 0$

$$\frac{\partial}{\partial \theta} \alpha_1 \ln \theta + \frac{\partial}{\partial \theta} \alpha_0 \ln (1 - \theta)$$

$$= \alpha_1 \frac{1}{\theta} + \alpha_0 \frac{\partial}{\partial \theta} \ln (1 - \theta)$$

$$= \frac{\alpha_1}{\theta} + \alpha_0 \frac{\partial \ln (1 - \theta)}{\partial (1 - \theta)} \frac{\partial (1 - \theta)}{\partial \theta} \quad [\text{chain Rule}] \quad \frac{\alpha_1}{\theta} - \alpha_0 \frac{1}{(1 - \theta)}$$

$$= \frac{\alpha_1}{\theta} + \alpha_0 \frac{1}{1 - \theta} (-1) = \frac{\alpha_1}{\theta} - \alpha_0 \frac{1}{(1 - \theta)} \quad \left[ \hat{\theta}_{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0} \right]$$

$$\arg \max_x [x + y + xy]$$

$$= \arg \max_x [x + xy]$$

$$\frac{\partial}{\partial \theta} \ln \theta = \frac{1}{\theta}$$



# MLE

- $X$  is  $\sim$  Bernoulli

- $P(X) = \theta^X (1 - \theta)^{(1-X)}$

- Likelihood is  $\sim$  Binomial

- $P(D | \theta) = \theta^{\alpha_1} (1 - \theta)^{\alpha_0}$

- $\hat{\theta}_{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$

# How many flips do we need?

$$\hat{\theta}_{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

$$\alpha_1 = 3, \quad \alpha_0 = 2$$

$$\hat{\theta}_{MLE} = \frac{3}{5} = 0.6$$

$$\alpha_1 = 51, \quad \alpha_0 = 49$$

$$\hat{\theta}_{MLE} = 0.51$$

# Bayesian thinking...

- Use Bayes Rule:

$$P(\theta | data) = \frac{P(data | \theta) P(\theta)}{P(data)}$$

MLE  
likelihood  
prior

posterior  
MAP

$$P(\theta | data) \propto P(Data | \theta) P(\theta)$$

# Bayesian thinking...

- Use Bayes Rule:

likelihood

posterior

prior

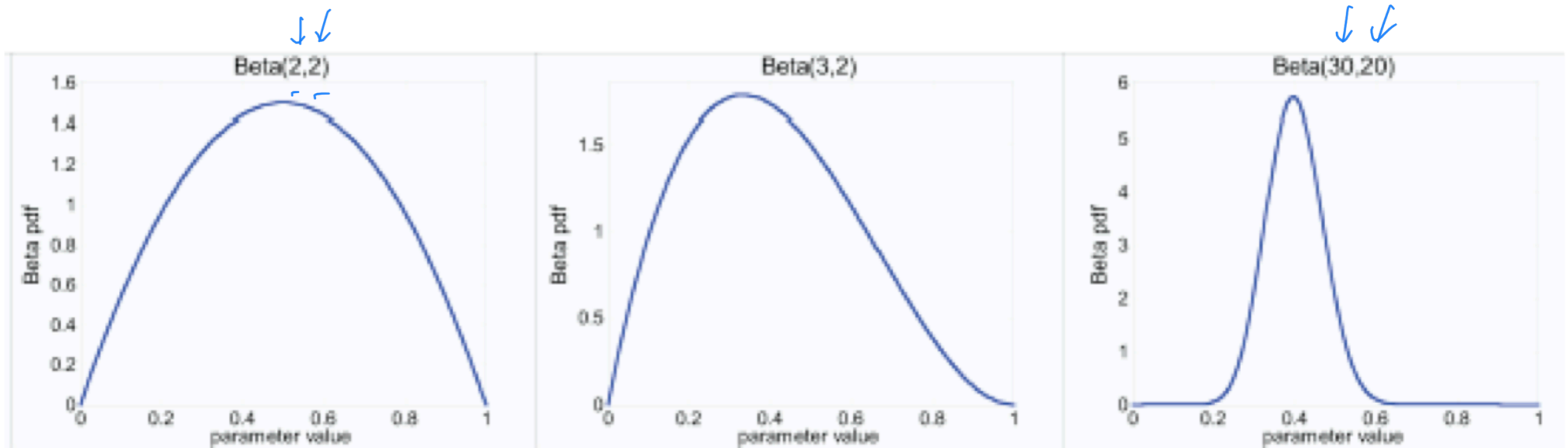
$$P(\theta | data) = \frac{P(data | \theta)P(\theta)}{P(data)}$$

- Equivalently,

$$P(\theta | data) \propto P(data | \theta)P(\theta)$$

# Beta prior distribution - $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}}{B(\beta_1, \beta_0)} \sim \text{Beta}(\beta_1, \beta_0)$$



# Beta prior distribution - $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_1-1} (1-\theta)^{\beta_0-1}}{B(\beta_1, \beta_0)} \sim \text{Beta}(\beta_1, \beta_0)$$

$$\text{Likelihood function: } P(D | \theta) = \theta^{\alpha_1} (1-\theta)^{\alpha_0}$$

$\beta_1 - 1 = \# \text{ balls in red heads}$   
 $\beta_0 - 1 = \# \text{ balls in red tails}$

$$\text{Posterior: } P(\theta | \text{data}) \propto P(\text{data} | \theta) P(\theta)$$

$\text{argmax}_{\theta} P(\theta | \text{data})$

$$\frac{\theta^{\alpha_1} (1-\theta)^{\alpha_0} \theta^{\beta_1-1} (1-\theta)^{\beta_0-1}}{B(\beta_1, \beta_0)}$$

$\text{argmax}_{\theta} \theta^{(\alpha_1+\beta_1-1)} (1-\theta)^{(\alpha_0+\beta_0-1)}$

$$= \frac{\theta^{(\alpha_1+\beta_1-1)} (1-\theta)^{(\alpha_0+\beta_0-1)}}{B(\beta_1, \beta_0)}$$

$$\hat{\theta}_{\text{MAP}} = \frac{(\alpha_1 + \beta_1 - 1)}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}$$



# Posterior distribution

- Prior:  $Beta(\beta_1, \beta_0)$
- Data :  $\alpha_1$  heads and  $\alpha_0$  tails
- Posterior distribution:  $P(\theta | data) \sim Beta(\beta_1 + \alpha_1, \beta_0 + \alpha_0)$

# MAP

$$\circ P(\theta | D) = \frac{\theta^{\beta_1 + \alpha_1 - 1} (1 - \theta)^{\beta_0 + \alpha_0 - 1}}{B(\beta_1 + \alpha_1, \beta_0 + \alpha_0)} \sim \text{Beta}(\beta_1 + \alpha_1, \beta_0 + \alpha_0)$$

$$\circ \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta | D) = \frac{(\alpha_1 + \beta_1)}{(\alpha_1 + \beta_1) + (\alpha_0 + \beta_0)}$$

- As  $N \rightarrow \infty$ , prior is “forgotten”
- But for small sample size, prior is important

# Conjugate prior

- $P(\theta)$  and  $P(D | \theta)$  have the same form

- Likelihood is  $\sim$  Binomial

- $P(D | \theta) = \theta^{\alpha_1} (1 - \theta)^{\alpha_0}$

- If prior is Beta distribution,

- $P(\theta) = \frac{\theta^{\beta_1 - 1} (1 - \theta)^{\beta_0 - 1}}{B(\beta_1, \beta_0)} \sim \text{Beta}(\beta_1, \beta_0)$

- Then posterior is Beta distribution

- $P(\theta | D) = \frac{\theta^{\beta_1 + \alpha_1 - 1} (1 - \theta)^{\beta_0 + \alpha_0 - 1}}{B(\beta_1 + \alpha_1, \beta_0 + \alpha_0)} \sim \text{Beta}(\beta_1 + \alpha_1, \beta_0 + \alpha_0)$

**For Binomial, conjugate prior is Beta distribution**

# Conjugate prior

- $P(\theta)$  and  $P(D | \theta)$  have the same form
- Dice roll problem (6 outcomes instead of 2)
- Likelihood is  $\sim$  Multinomial  $\theta = \{\theta_1, \theta_2, \dots, \theta_k\}$

- $P(D | \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \dots \theta_k^{\alpha_k}$

- If prior is Dirichlet distribution,

- $P(\theta) = \frac{\prod_{i=1}^k \theta_i^{\beta_i - 1}}{B(\beta_1, \dots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \dots, \beta_k)$

- Then posterior is Dirichlet distribution
- $P(\theta | D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \dots, \beta_k + \alpha_k)$

**For Multinomial, conjugate prior is Dirichlet distribution**