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

- **Principle 1.** Maximum likelihood estimation (**MLE**)
  - Choose parameter θ that maximizes $P(data \mid \theta)$
  - $\hat{\theta} = \dfrac{\alpha_1}{\alpha_1 + \alpha_0}$
- **Principle 2.** Maximum a posteriori estimation (**MAP**)
  - Choose parameter θ that maximizes $P(\theta \mid data)$
  - $\hat{\theta} = \dfrac{(\alpha_1 + \beta_1)}{(\alpha_1 + \beta_1) + (\alpha_0 + \beta_0)}$

Formal treatment

- **Principle 1.** Maximum likelihood estimation (**MLE**)
  - $P(X=1)=\theta\,; P(X=0)=1-\theta$
  - Data D : 1 0 0 1 1
  - $P(D\,|\,\theta)$: θ $(1-\theta)$ $(1-\theta)$ θ θ = $\theta^{\alpha_1}(1-\theta)^{\alpha_0}$

Flip produces data D with α_1 heads and α_0 tails, iid ~ Bernoulli

α_1 heads and α_0 are counts that sum these outcomes ~ Binomial

Learning θ 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

- $P(D \mid \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$
- $\hat{\theta} = \arg\max_{\theta} P(D \mid \theta) = \arg\max_{\theta} \ln P(D \mid \theta)$
  - Set the derivative to zero $\frac{d}{d\theta} \ln P(D \mid \theta) = 0$

MLE

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

- Likelihood is ~ Binomial
  - $P(D \mid \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$

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

How many flips do we need?

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

Bayesian thinking…

- Use Bayes Rule:
  - $$P(\theta \mid data) = \frac{P(data \mid \theta)P(\theta)}{P(data)}$$

Bayesian thinking…

- Use Bayes Rule:

likelihood

posterior

prior

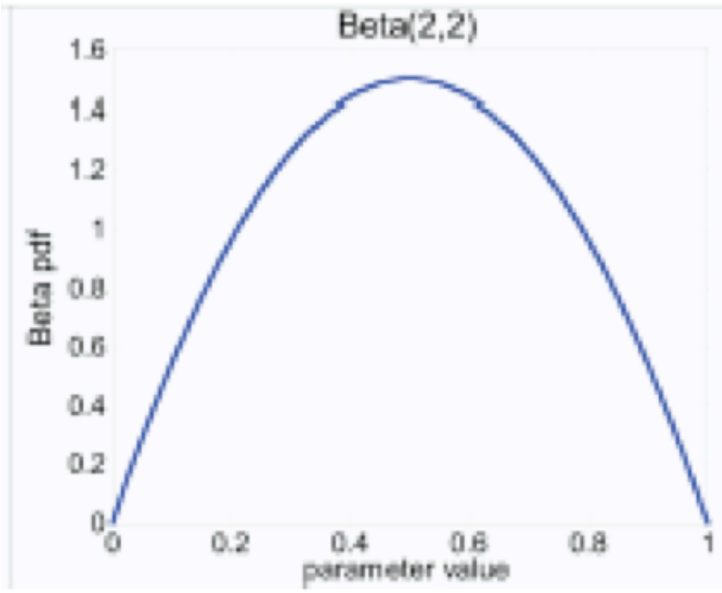
$$P(\theta \mid data) = \frac{P(data \mid \theta)P(\theta)}{P(data)}$$

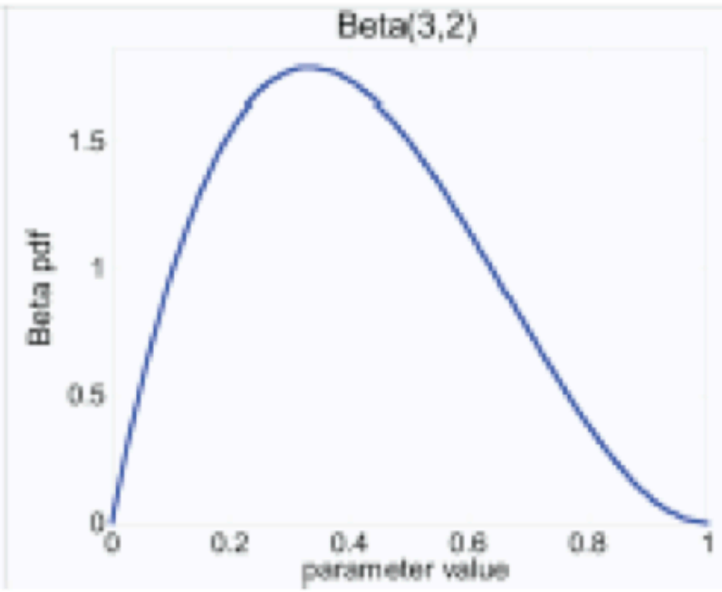
- Equivalently,

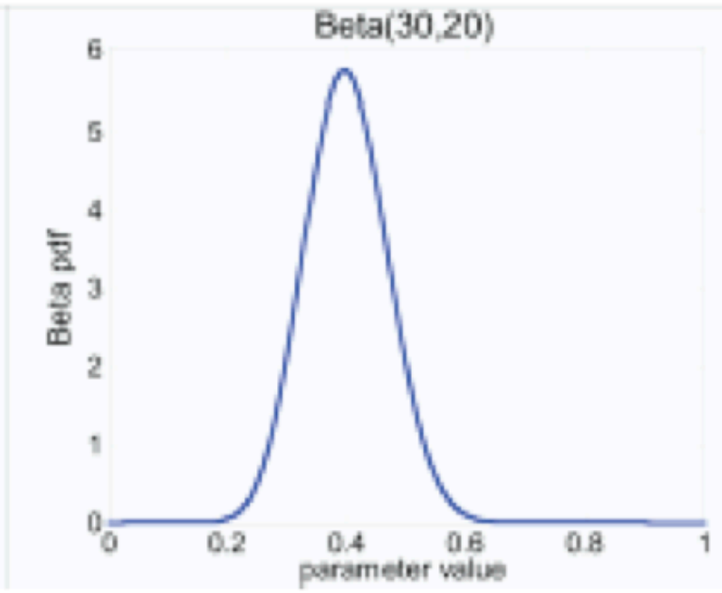
$$P(\theta \mid data) \propto P(data \mid \theta)P(\theta)$$

Beta prior distribution - $P(\theta)$

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







Beta prior distribution - $P(\theta)$

- $P(\theta) = \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{B(\beta_1, \beta_0)} \sim Beta(\beta_1, \beta_0)$
- Likelihood function: $P(D \mid \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$
- Posterior: $P(\theta \mid data) \propto P(data \mid \theta)P(\theta)$

Posterior distribution

- Prior: $Beta(\beta_1, \beta_0)$

- Data : α_1 heads and α_0 tails

- Posterior distribution: $P(\theta \mid data) \sim Beta(\beta_1 + \alpha_1, \beta_0 + \alpha_0)$

MAP

- $P(\theta \mid D) = \dfrac{\theta^{\beta_1+\alpha_1-1}(1-\theta)^{\beta_0+\alpha_0-1}}{B(\beta_1+\alpha_1, \beta_0+\alpha_0)} \sim Beta(\beta_1+\alpha_1, \beta_0+\alpha_0)$
- $\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \dfrac{(\alpha_1+\beta_1)}{(\alpha_1+\beta_1)+(\alpha_0+\beta_0)}$

- As $N \to \infty$, 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 ~ Binomial
  - $P(D|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$
- If prior is Beta distribution,
  - $P(\theta) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}}{B(\beta_1,\beta_0)} \sim Beta(\beta_1,\beta_0)$
- Then posterior is Beta distribution
  - $P(\theta|D) = \frac{\theta^{\beta_1+\alpha_1-1}(1-\theta)^{\beta_0+\alpha_0-1}}{B(\beta_1+\alpha_1,\beta_0+\alpha_0)} \sim Beta(\beta_1+\alpha_1,\beta_0+\alpha_0)$

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 ~ Multinomial $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$
  - $P(D|\theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2}\ldots\theta_k^{\alpha_k}$

- If prior is Dirichlet distribution,
  - $P(\theta) = \frac{\prod_{i=1}^{k}\theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim Dirichlet(\beta_1, \ldots \beta_k)$

- Then posterior is Dirichlet distribution

- $P(\theta|D) \sim Dirichlet(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$

For Multinomial, conjugate prior is Dirichlet distribution