CS440 Introduction to Artificial Intelligence

Lecture 20: Bayesian learning; conjugate priors

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

Office Hours: Thu, 2:00-3:00pm

http://www.cs.uiuc.edu/class/sp11/cs440

The binomial distribution

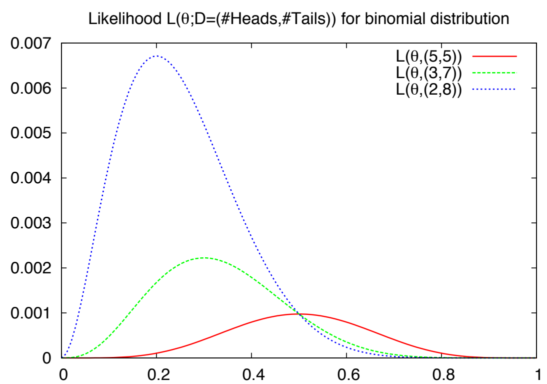
If p is the probability of heads, the probability of getting exactly k heads in n independent yes/no trials is given by the binomial distribution $Bin(n,p)$:

$$P(k \text{ heads}) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Expectation $E(Bin(n,p)) = np$
Variance $var(Bin(n,p)) = np(1-p)$

Binomial likelihood

What distribution does p (probability of heads) have, given that the data D consists of #H heads and #T tails?



Parameter estimation

Given data D=$HTTHTT$, what is the probability θ of heads?

- **Maximum likelihood estimation (MLE):**
  Use the θ which has the **highest likelihood** $P(D|\theta)$.
  $\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$
- **Maximum a posterior (MAP):**
  Use the θ which has the **highest posterior probability** $P(\theta|D)$.
  $\theta_{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(\theta)P(D|\theta)$
- **Bayesian estimation:**
  Integrate over all θ = compute the **expectation of θ given D**:
  $P(x=H|D) = \int_0^1 P(x=H|\theta)P(\theta|D)d\theta = E[\theta|D]$

Maximum likelihood estimation

- **Maximum likelihood estimation (MLE):**
  find θ which maximizes likelihood $P(D \mid \theta)$.

$$\begin{aligned}\theta^* &= \arg\max_\theta P(D|\theta)\\ &= \arg\max_\theta \theta^H(1-\theta)^T\\ &= \frac{H}{H+T}\end{aligned}$$

Bayesian statistics

- Data D provides evidence for or against our beliefs.
  We update our belief θ based on the evidence we see:

$$\underbrace{P(\theta|D)}_{\text{Posterior}} = \frac{\overbrace{P(\theta)}^{\text{Prior}}\overbrace{P(D|\theta)}^{\text{Likelihood}}}{\underbrace{\int P(\theta)P(D|\theta)d\theta}_{\text{Marginal Likelihood } (=P(D))}}$$

Bayesian estimation

**Given a prior $P(\theta)$ and a likelihood $P(D|\theta)$,
what is the posterior $P(\theta|D)$?**

How do we choose the prior $P(\theta)$?

- The posterior is proportional to prior x likelihood:
  $P(\theta|D) \propto P(\theta)\, P(D|\theta)$
- The likelihood of a binomial is:
  $P(D|\theta) = \theta^H(1-\theta)^T$
- If prior $P(\theta)$ is proportional to powers of θ and $(1-\theta)$,
  posterior will also be proportional to powers of θ and $(1-\theta)$:
  $P(\theta) \propto \theta^a(1-\theta)^b$
  $\Rightarrow P(\theta|D) \propto \theta^a(1-\theta)^b\,\theta^H(1-\theta)^T = \theta^{a+H}(1-\theta)^{b+T}$

In search of a prior...

We would like something of the form:

$$P(\theta) \propto \theta^a(1-\theta)^b$$

But -- this looks just like the binomial:

$$\begin{aligned}P(k \text{ heads}) &= \binom{n}{k}p^k(1-p)^{n-k}\\ &= \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\end{aligned}$$

.... except that k is an integer and θ is a real with $0<\theta<1$.

The Gamma function

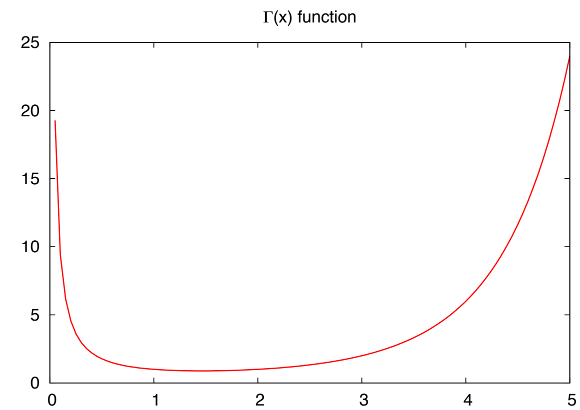
The Gamma function $\Gamma(x)$ is the generalization of the factorial $x!$ (or rather $(x-1)!$) to the reals:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad \text{for } \alpha > 0$$

For $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$.

For positive integers, $\Gamma(x) = (x-1)!$

The Gamma function



The Beta distribution

A random variable X ($0 < x < 1$) has a Beta distribution with (hyper)parameters α ($\alpha > 0$) and β ($\beta > 0$) if X has a continuous distribution with probability density function

$$P(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

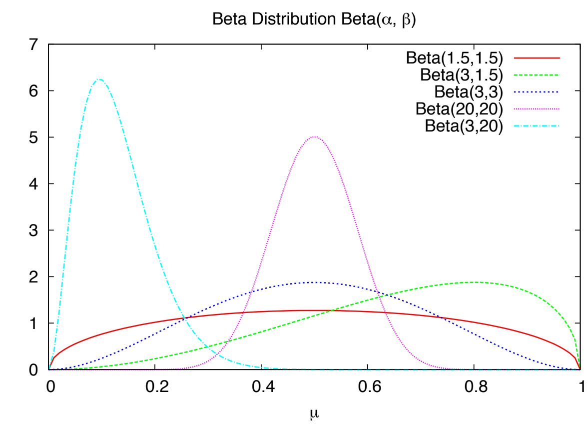
The first term is a normalization factor (to obtain a distribution)

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Expectation: $\frac{\alpha}{\alpha+\beta}$

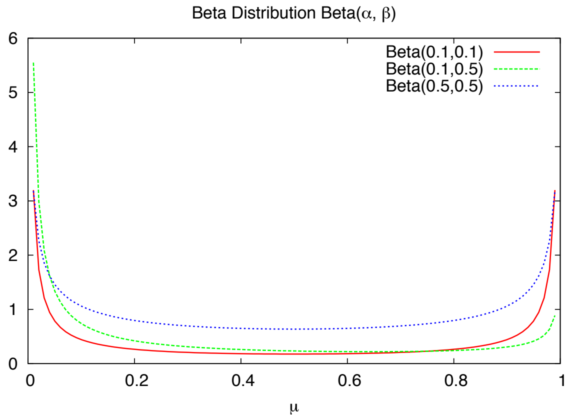
Beta(α,β) with *α >1, β >1*

Unimodal



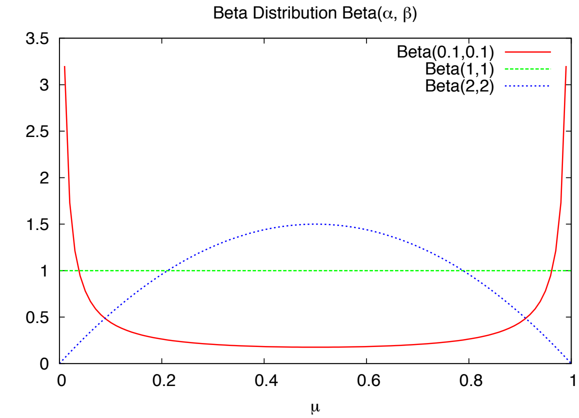
Beta(α,β) with *α <1, β <1*

U-shaped



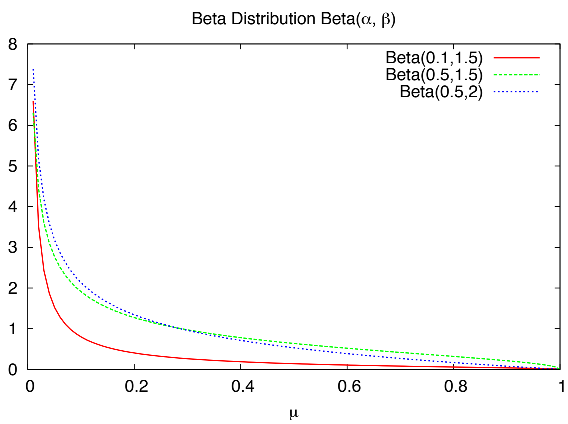
Beta(α,β) with *α=β*

Symmetric. $\alpha=\beta=1$: uniform



Beta(α,β) with *α<1, β >1*

Strictly decreasing

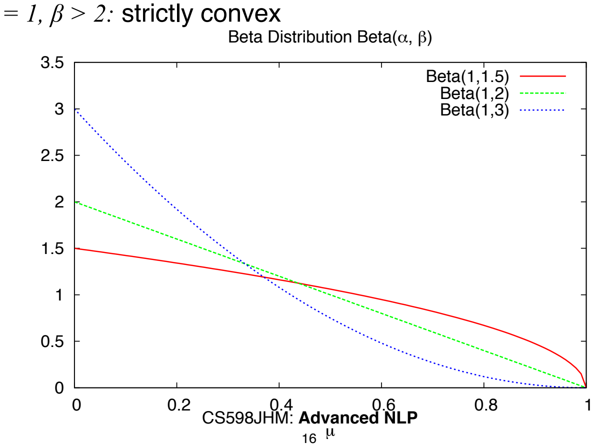


Beta(α,β) with *α = 1, β >1*

$\alpha = 1$, $1 < \beta < 2$: strictly concave.

$\alpha = 1$, $\beta = 2$: straight line

$\alpha = 1$, $\beta > 2$: strictly convex



Beta as prior for binomial

Given a **prior** $P(\theta\,|\alpha,\beta) = \text{Beta}(\alpha,\beta)$, and **data** $D=(H,T)$, what is our posterior?

$$\begin{aligned} P(\theta|\alpha,\beta,H,T) &\propto P(H,T|\theta)P(\theta|\alpha,\beta) \\ &\propto \theta^H(1-\theta)^T\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{H+\alpha-1}(1-\theta)^{T+\beta-1} \end{aligned}$$

With normalization

$$\begin{aligned} P(\theta|\alpha,\beta,H,T) &= \frac{\Gamma(H+\alpha+T+\beta)}{\Gamma(H+\alpha)\Gamma(T+\beta)}\theta^{H+\alpha-1}(1-\theta)^{T+\beta-1} \\ &= \text{Beta}(\alpha+H,\beta+T) \end{aligned}$$

So, what do we predict?

Our Bayesian estimate for the next coin flip $P(x=1 \mid D)$:

$$\begin{aligned} P(x=H|D) &= \int_0^1 P(x=H|\theta)P(\theta|D)d\theta \\ &= \int_0^1 \theta P(\theta|D)d\theta \\ &= E[\theta|D] \\ &= E[Beta(H+\alpha,T+\beta)] \\ &= \frac{H+\alpha}{H+\alpha+T+\beta} \end{aligned}$$

Conjugate priors

The beta distribution is a **conjugate prior** to the binomial: the resulting posterior is also a beta distribution.

We can interpret its parameters α, β as pseudocounts
$P(H \mid D) = (H+\alpha)/(H+\alpha+T+\beta)$

All members of the *exponential family* of distributions have conjugate priors.

Examples:
- Multinomial: conjugate prior = Dirichlet
- Gaussian: conjugate prior = Gaussian

Multinomials: Dirichlet prior

Multinomial distribution:
Probability of observing each possible outcome c_i exactly X_i times in a sequence of n yes/no trials:

$$P(X_1=x_i,\ldots,X_K=x_k) = \frac{n!}{x_1!\cdots x_K!}\theta_1^{x_1}\cdots\theta_K^{x_K} \quad \text{if} \quad \sum_{i=1}^{N} x_i = n$$

Dirichlet prior:

$$Dir(\theta|\alpha_1,\ldots\alpha_k) = \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\prod_{k=1}\theta_k^{\alpha_k-1}$$

More about conjugate priors

- We can interpret the hyperparameters as “pseudocounts”
- Sequential estimation (updating counts after each observation) gives same results as batch estimation
- Add-one smoothing (Laplace smoothing) = uniform prior
- On average, more data leads to a sharper posterior (sharper = lower variance)