CS598JHM: **Advanced NLP** (Spring '10)

Lecture 3

Julia Hockenmaier

juliahmr@illinois.edu

3324 Siebel Center

http://www.cs.uiuc.edu/class/sp10/cs598jhm

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| θ).
  $\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$
- **Maximum a posterior (MAP):**
  Use the θ which has the **highest posterior probability** P(θ |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**:
  $$P(x = H|D) = \int_0^1 P(x = H|\theta)P(\theta|D)d\theta = E[\theta|D]$$

Conjugate priors

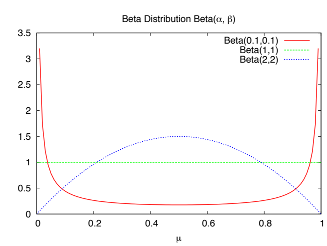
The **posterior** is proportional to **prior** x **likelihood**:
$P(\theta\,|D) \propto P(\theta)\, P(D|\theta)$

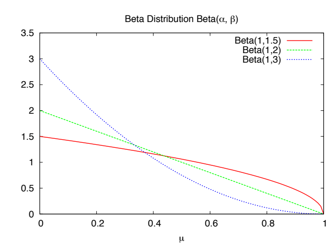
Conjugate priors:
Posterior is the same kind of distribution as prior.

For **binomial likelihood:**
conjugate prior = **Beta distribution**

The Beta distribution

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





Beta as prior for binomial

Posterior for prior $P(\theta\,|\alpha,\beta) = \mathrm{Beta}(\alpha,\beta)$, and data $D=(H,T)$:

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

$$P(\theta|\alpha,\beta,H,T) = \mathrm{Beta}(\alpha+H,\beta+T)$$

Prediction for next coin flip:

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

Multinomial variables

- In NLP, X is often a **discrete** random variable that can take one of K states.
- We can represent such Xs as **K-dimensional vectors** in which one $x_k = 1$ and all other elements are 0
  $x = (0,0,1,0,0)^T$
- Denote probability of $x_k = 1$ as μ_k with $0 \leq \mu_k \leq 1$ and $\sum_k \mu_k = 1$
  Then the probability of x is:

$$P(x|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

Multinomial likelihood

- What is the **likelihood** of $D = x_1 \dots x_i \dots x_N$?

$$\begin{aligned} P(D|\boldsymbol{\mu}) &= \prod_{i=1}^{N} P(x_i|\boldsymbol{\mu}) \\ &= \prod_{i=1}^{N}\prod_{k=1}^{K} \mu_k^{x_{nk}} \\ &= \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} \\ &:= \prod_{k=1}^{K} \mu_k^{m_k} \end{aligned}$$

Define

$$m_k := \sum_{n=1}^{N} x_{nk}$$

(= #observations of xk=1)

The likelihood depends only on the m_ks.
m_k are **sufficient statistics**

Multinomials: Dirichlet prior

The joint distribution of $(m_1, \dots, m_K)$ conditioned on $\boldsymbol{\mu}$ and N is a **multinomial distribution:**

$$P(m_1,\dots,m_K = x_k) = \frac{N!}{m_1!\cdots m_K!}\theta_1^{m_1}\cdots\theta_K^{x_K}$$

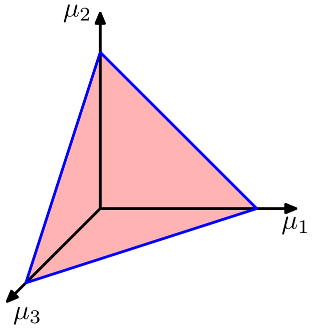
$$\text{if } \sum_{i=1}^{K} x_k = N$$

Multinomials have a **Dirichlet prior:**

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

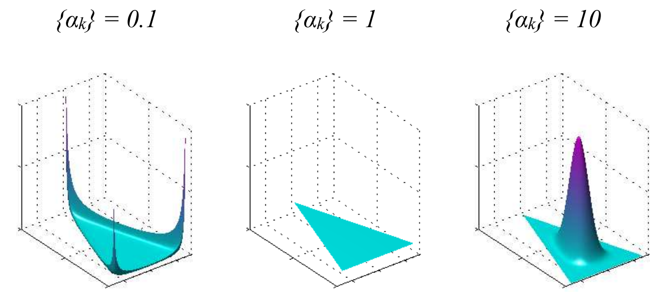
The Dirichlet

A Dirichlet is confined to a simplex (here $\boldsymbol{\mu}=(\mu_1,\mu_2,\mu_3)$)



(Figure from Chris Bishop's PRML book & website)

Examples of the Dirichlet



(all figures from Chris Bishop's PRML book & website)

Dirichlet as conjugate prior

Given a prior $Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})$ and Data D with sufficient statistics $\boldsymbol{m}=(m_1,...,m_K)$, the posterior is

$$\begin{aligned} p(\boldsymbol{\mu}|D,\boldsymbol{\alpha}) &\propto P(D|\boldsymbol{\mu})P(\boldsymbol{\mu}) \\ &\propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1 + m_k} \end{aligned}$$

The normalized posterior is:

$$\begin{aligned} p(\boldsymbol{\mu}|D,\boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_1 + ... + \alpha_K + N)}{\Gamma\alpha_1 + m_1 ... \Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1 + m_k} \end{aligned}$$

MLE vs Bayesian estimate

Maximum likelihood estimate:
Maximize $\ln p(D|\boldsymbol{\mu})$ wrt. μ_k under the constraint that $\sum \mu_k = 1$

(...Use Lagrange multipliers…)

$$\mu_k^{MLE} = \frac{m_k}{N}$$

Bayesian estimate:

$$\mu_k^{BE} = \frac{m_k + \alpha_k}{N + \sum_{k'} \alpha'_k}$$