

# CS281A/Stat241A recitation 5: Priors, posteriors, MLE

October 1, 2007

Percy Liang

## Introduction

- Bayesian inference
  - Have a likelihood model  $p(x | \theta)$ , where data  $x = (x_1, \dots, x_n)$  is i.i.d.
  - Use a conjugate prior  $p(\theta)$
  - Use Bayes rule to derive posterior  $p(\theta | x) = \frac{p(\theta)p(x|\theta)}{p(x)} \propto p(\theta)p(x | \theta)$
  - Strategy:
    - \* Multiply  $p(\theta)$  and  $p(x | \theta)$
    - \* Can drop constants (factors or terms in exponent that don't depend on  $\theta$ )
    - \* Recognize the normalization constant immediately
- Maximum likelihood
  - $\theta^* = \operatorname{argmax}_{\theta} p(x | \theta)$
  - Strategy:
    - \* Form the Lagrangian (if  $\theta$  has constraints)
    - \* Differentiate  $\frac{\partial}{\partial \theta} \log p(x | \theta)$
    - \* Set it to zero and solve for  $\theta$ , using the constraints
- Relationship between the two: if  $p(\theta) \propto 1$  (the prior is “uniform”) then the maximum likelihood estimate is the mode of the posteriors ( $\theta^* = \operatorname{argmax}_{\theta} p(\theta | x)$ )

## Gaussian mean parameter

- Likelihood model:  $\theta = \mu$

$$p(x | \mu) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

- Conjugate prior is Gaussian:

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left\{ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right\}$$

- Work with precisions  $\tau = 1/\sigma^2$  and  $\tau_0 = 1/\sigma_0^2$  rather than variances because the results are cleaner and more intuitive
- Multiply to get posterior (up to a constant):

$$\begin{aligned} p(\mu | x) &\propto \exp \left\{ -\frac{1}{2} (\tau(\mu - x)^2 + \tau_0(\mu - \mu_0)^2) \right\} \\ &= \exp \left\{ -\frac{1}{2} ((\tau + \tau_0)\mu^2 + 2(\tau x + \tau_0\mu_0)\mu) \right\} \end{aligned}$$

– Define

- \*  $\tau_1 \stackrel{\text{def}}{=} \tau + \tau_0$ : new precision is sum of prior precision and likelihood precision
- \*  $\mu_1 = \frac{\tau x + \tau_0 \mu_0}{\tau + \tau_0}$ : new mean is convex combination of prior mean and data

– Completing the square:  $\tau_1 \mu^2 - 2\tau_1 \mu_1 \mu = \tau_1 (\mu - \mu_1)^2 + \text{constant}$

$$p(\mu | x) \propto \exp \left\{ -\frac{1}{2} \tau_1 (\mu - \mu_1)^2 \right\}$$

Define  $\frac{1}{\sigma_1^2} = \tau_1$  and identify  $p(\mu | x) = \mathcal{N}(\mu; \mu_1, \sigma_1^2)$ .

- Multiple data points: Let  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ . Trick:

$$\begin{aligned} \sum_i (x_i - \mu)^2 &= \sum_i [(x_i - \bar{x}) + (\bar{x} - \mu)]^2 \\ &= \sum_i [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2] \\ &= \left( \sum_i (x_i - \bar{x})^2 \right) + 2(\bar{x} - \mu) \underbrace{\sum_i (x_i - \bar{x})}_{=0} + n(\bar{x} - \mu)^2 \\ &= \underbrace{\left( \sum_i (x_i - \bar{x})^2 \right)}_{\text{constant}} + n(\bar{x} - \mu)^2 \end{aligned}$$

$$p(x_1, \dots, x_n | \mu) \propto \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \propto \exp \left\{ -\frac{n\tau}{2} (\bar{x} - \mu)^2 \right\}$$

so we can use machinery for one data point, replacing  $x$  with  $\bar{x}$  and  $\tau$  with  $n\tau$ .

- Another view: note that  $p(\theta | x_1, \dots, x_n) \propto p(\theta | x_1, \dots, x_{n-1})p(x_n | \theta)$ , so we can incorporate the points into the posterior sequentially.
- Maximum likelihood:
  - See first recitation or book for direct derivation.
  - Mean/mode is  $\frac{\tau_0 \mu_0 + n\tau \bar{x}}{\tau_0 + n\tau} \rightarrow \bar{x}$  as  $\tau_0 \rightarrow 0$  (weak prior) or  $n \rightarrow \infty$  (more data)

## Gaussian variance parameter

- Likelihood model:  $\theta = \sigma^2$

$$p(x | \sigma^2) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} \propto (\sigma^2)^{-1/2} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

- Conjugate prior is Inverse gamma

$$p(\sigma^2) = \text{IG}(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp\left\{\frac{-\beta}{\sigma^2}\right\}$$

- Compute posterior:

$$\begin{aligned} p(\sigma^2 | x) &\propto p(\sigma^2)p(x | \sigma^2) \\ &\propto (\sigma^2)^{-(\alpha+1/2)-1} \exp\left\{-\frac{\beta + \frac{1}{2}(x - \mu)^2}{2\sigma^2}\right\} \\ &\propto \text{IG}\left(\sigma^2; \alpha + \frac{1}{2}, \beta + \frac{1}{2}(x - \mu)^2\right) \end{aligned}$$

- Maximum likelihood:

– See first recitation or book for direct derivation.

– Inverse Gamma prior on  $\sigma^2$

\* Mean:  $\frac{\beta + \frac{1}{2} \sum_i (x_i - \mu)^2}{\alpha + \frac{n}{2} - 1}$

\* Mode:  $\frac{\beta + \frac{1}{2} \sum_i (x_i - \mu)^2}{\alpha + \frac{n}{2} + 1} = \frac{1}{n+2} \sum_i (x_i - \mu)^2$  with  $\alpha = \beta = 0$

\* Flat prior decreases variance estimate

– Gamma( $\alpha, \beta$ ) prior on precision  $\tau$

\* Mode:  $\frac{\alpha + \frac{n}{2}}{\beta + \frac{1}{2} \sum_i (x_i - \mu)^2}$

\* Mode:  $\frac{\alpha + \frac{n}{2} - 1}{\beta + \frac{1}{2} \sum_i (x_i - \mu)^2} = \left(\frac{1}{n-2} \sum_i (x_i - \mu)^2\right)^{-1}$  with  $\alpha = \beta = 0$

\* Flat prior increases variance estimate

## Multinomial parameter

- $x$  is an indicator vector (all zeros except for exactly one index)

- Likelihood:  $p(x | \theta) = \prod_{j=1}^d \theta_j^{x_j}$

- Conjugate prior is Dirichlet

$$p(\theta) = \text{Dir}(\theta; \alpha) = \frac{\Gamma(\sum_{j=1}^d \alpha_j)}{\underbrace{\prod_{j=1}^d \Gamma(\alpha_j)}_{\frac{1}{D(\alpha)}}} \prod_{j=1}^d \theta_j^{\alpha_j - 1}$$

- Compute posterior:

$$p(\theta | x) \propto \prod_{j=1}^d \theta_j^{\alpha_j + x_j - 1} \propto \text{Dir}(\theta; \alpha + x)$$

- Marginal likelihood:

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

- Maximum likelihood via posterior:

- Mean:  $\theta_j = \frac{\alpha_j}{\sum_{j'=1}^d \alpha_{j'}}$
- Mode:  $\theta_j = \frac{\alpha_j - 1}{\sum_{j'=1}^d (\alpha_{j'} - 1)}$

- Maximum likelihood directly via Lagrange multipliers:

- Objective:

$$\max_{\theta} \log p(x | \theta) = \max_{\theta} \sum_{j=1}^d x_j \log \theta_j$$

- Constraints:

$$\sum_{j=1}^d \theta_j = 1 \quad [\text{introduce Lagrange multipliers}]$$

$$\theta_j \geq 0 \text{ for all } j \quad [\text{handle explicitly}]$$

- Form the Lagrangian

$$L(\theta, \lambda) = \sum_{j=1}^d x_j \log \theta_j - \lambda \left( \sum_{j=1}^d \theta_j - 1 \right)$$

- Optimal solution must satisfy

- \* For all  $j = 1, \dots, d$ :

$$\frac{\partial}{\partial \theta_j} L(\theta, \lambda) = \frac{x_j}{\theta_j} - \lambda = 0$$

$$\text{so } \theta_j = \frac{x_j}{\lambda}.$$

- \*  $\sum_j \theta_j = 1$ , so  $\theta_j = \frac{x_j}{\sum_{j'} x_{j'}}$