Parameter Estimation II: Bayesian Parameter Estiimation

Matthew Stone
CS 520, Spring 2000
Lecture 5

Parameter Estimation as Learning

- **Use training data to estimate unknown probabilities and probability density functions**

Framework for Combining Information

- **Bayesian estimation**
  - We have expectations about parameters
    - Expectations are expressed as prior density on parameter values
    - We want to combine these expectations with measurements (training data)
- **Use Bayes's formula to derive a posterior**
  - First for parameter values
  - Then for future measurements

More Precisely

(using notation from last time)

- **Want to derive**

  density $p(\mathbf{x} \,|\, \mathcal{D})$ from samples $\mathcal{D}$
- **When**
  - Form of density $p(\mathbf{x}|\theta)$ is known
  - Initial knowledge gives prior density $p(\theta)$
  - Remaining knowledge comes from n samples drawn independently by $p(\mathbf{x}|\theta)$

Approaching the Density

- **Reason by cases in parameter space**

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x}, \theta \mid \mathcal{D}) d\theta$$

- **Factor derivation of x through θ:**

$$\begin{aligned} p(\mathbf{x}, \theta \mid \mathcal{D}) &= p(\mathbf{x} \mid \theta, \mathcal{D}) p(\theta \mid \mathcal{D}) \\ &= p(\mathbf{x} \mid \theta) p(\theta \mid \mathcal{D}) \end{aligned}$$

$$\boxed{p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x} \mid \theta) p(\theta \mid \mathcal{D}) d\theta}$$

Approaching the Density, II

- **Get the posterior on parameters by Bayes**

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) p(\theta)}{\int p(\mathcal{D} \mid \theta) p(\theta) d\theta}$$

A Bit of Magic

- **What if we get a new piece of data?**

$$p(\theta \mid \mathscr{D}, \mathbf{x}) = \frac{p(\mathscr{D}, \mathbf{x} \mid \theta)\, p(\theta)}{\int p(\mathscr{D}, \mathbf{x} \mid \theta)\, p(\theta)\, d\theta}$$

$$= \frac{p(\mathbf{x} \mid \theta)\, p(\theta \mid \mathscr{D})}{\int p(\mathbf{x} \mid \theta)\, p(\theta \mid \mathscr{D})\, d\theta}$$

In Other Words

- **We start with a prior *p*(θ)**
- **We get a point and compute a posterior**

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta)\, p(\theta)}{\int p(\mathbf{x} \mid \theta)\, p(\theta)\, d\theta}$$

- **We get another point and update that**

$$p(\theta \mid \mathbf{x}, \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)\, p(\theta \mid \mathbf{x})}{\int p(\mathbf{y} \mid \theta)\, p(\theta \mid \mathbf{x})\, d\theta}$$

And So On

(and so on and so on…)

- **Recursive parameter estimation**
  - Incremental (on-line) learning
  - Using all available information

Bayes vs. Maximum Likelihood

Extended Illustration

- **One-dimensional samples**
  - Uniform distribution, unknown range

$$p(x \mid \theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

  - Need to estimate θ from data

Bayes vs. Maximum Likelihood
Deriving MLE Estimate

- **Maximum likelihood**
  - Given *n* data points *D*
  - Estimate of θ is the maximum of *D*

$$p(D \mid \theta) \propto \begin{cases} 1/\theta^n & \theta \geq \max D \\ 0 & \text{otherwise} \end{cases}$$

Bayes vs. Maximum Likelihood
Plan for Bayes Estimation

- **Recursive Bayes learning**
  - Given uniform prior on θ

$$p(\theta) \sim U(0{,}10)$$

  - Derive a series of more precise estimates

$$p(\theta \mid D^i)$$
$$p(x \mid D^i)$$

Bayes vs. Maximum Likelihood
First Data Point – Parameters

- **Estimate for the first data point *d***
  - The formula we derived is:

$$p(\theta \mid d) = \frac{p(d \mid \theta)\, p(\theta)}{\int p(d \mid \theta)\, p(\theta)\, d\theta}$$

  - Here that means:

$$p(\theta \mid d) \propto p(d \mid \theta) = \begin{cases} 1/\theta & d \le \theta \le 10 \\ 0 & \text{otherwise} \end{cases}$$

Bayes vs. Maximum Likelihood
First Data Point – Measurements

- **Estimate for the first data point *d***
  - We get the posterior on measurements as:

$$\boxed{p(x \mid d) = \int p(x \mid \theta)\, p(\theta \mid d)\, d\theta}$$

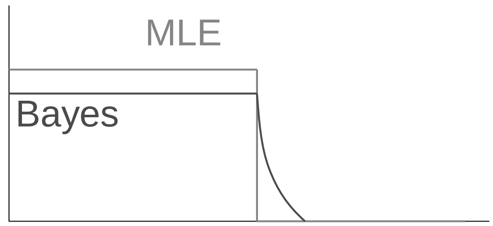
  - Here that means:

$$p(x \mid d) \propto \int_{d}^{10} p(x \mid \theta) \frac{1}{\theta}\, d\theta = \int_{\max\{x,d\}}^{10} \frac{1}{\theta^2}\, d\theta$$

$$p(x \mid d) \propto \begin{cases} 1/d - 1/10 & 0 \le x \le d \\ 1/x - 1/10 & d < x \le 10 \end{cases}$$

Bayes vs. Maximum Likelihood
Contrast, I

- **Different estimates for *p*(*x*|*d*)**



- **Bayes gives tail at higher values – prior info that this is possible balanced w data**

Bayes vs. Maximum Likelihood
Second Data Point – Parameters

- **Estimate for the next data point *d′***
  - The next stage of estimation is:

$$p(\theta \mid d, d') = \frac{p(d' \mid \theta)\, p(\theta \mid d)}{\int p(d' \mid \theta)\, p(\theta \mid d)\, d\theta}$$

  - Here that means:

$$p(\theta \mid d, d') \propto p(d' \mid \theta)\, p(\theta \mid d)$$

$$p(\theta \mid d, d') \propto \begin{cases} 1/\theta^2 & \max\{d, d'\} \le \theta \le 10 \\ 0 & \text{otherwise} \end{cases}$$