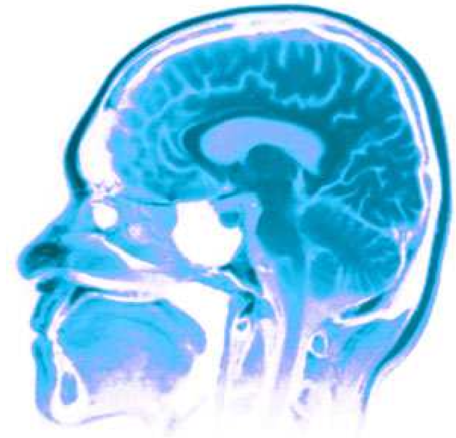
CPSC340



Bayesian learning

Nando de Freitas
October, 2012
University of British Columbia

Maximum likelihood revision

Find θ by maximizing $P(\text{data} \mid \theta)$.

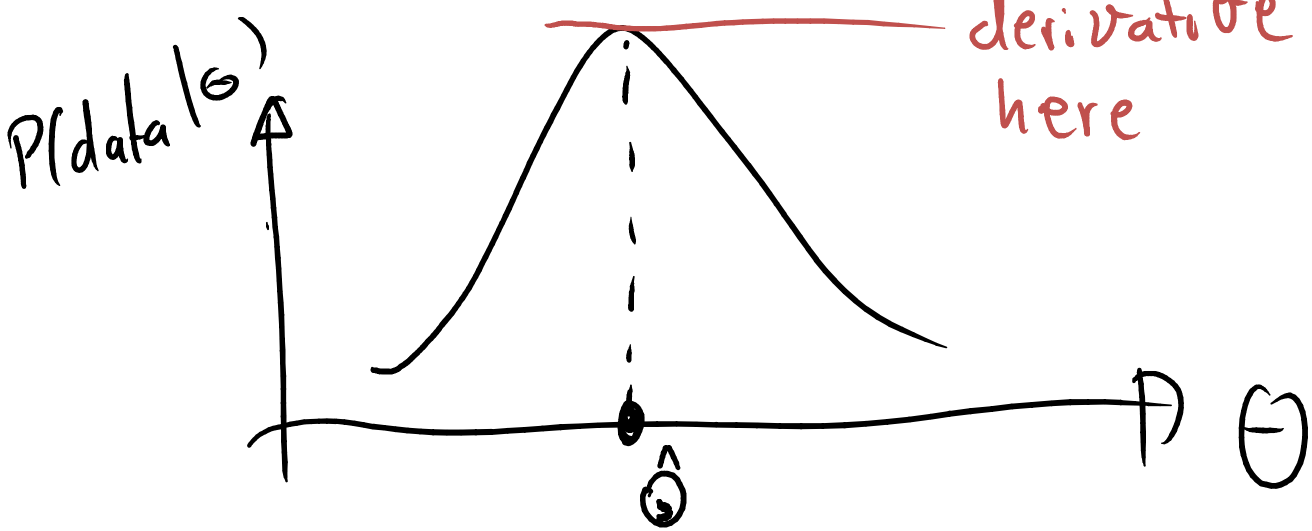
e.g. for a coin $P(\underbrace{\text{data}}_{x_{1:n}} \mid \theta) = \theta^m (1-\theta)^{n-m}$

$m = \#\ 1\text{'s}$

$n = \#\ 1\text{'s and } 0\text{'s}$

Slope

derivative is zero here



Maximum likelihood revision

$\hat{\Theta}_{ML}$ is the theta that minimizes

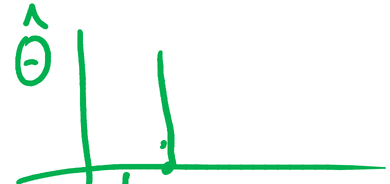
$$error(\theta) = Information(\Theta_0) - Information(\Theta)$$

true theta

any theta

Θ_0

$\hat{\Theta}$



$X = 111101100111$

$1100010010 = \tilde{X}$

choose $\hat{\Theta}$ to minimize $|f(x) - f(\tilde{x})|$

Outline of the lecture

This lecture introduces us to our second strategy for learning: **Bayesian learning**. The goal is for you to learn:

- ❑ Definition Beta prior.
- ❑ How to use Bayes rule to go from **prior beliefs** and the **likelihood of the data** to **posterior beliefs**.

Bayesian learning procedure

Step 1: *Given* **n** *data,* $x_{1:n} = \{x_1, x_2, \ldots, x_n\}$*, write down the expression for the likelihood:*

$$p(x_{1:n} \mid \theta) = \theta^m (1-\theta)^{n-m} \quad (\text{for a coin})$$

Step 2: *Specify a prior:* $p(\theta)$

$$P(\theta \mid x_{1:n}) = \underbrace{\frac{1}{\quad}}_{\text{const}} P(x_{1:n} \mid \theta) P(\theta)$$

Step 3: *Compute the posterior:*

$$\boxed{p(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)\, p(\theta)}{p(x_{1:n})}}$$

Notation:

$$p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta)\, p(\theta)$$

proportional to

Bayesian learning procedure

$P(A) = \sum_B P(AB)$

$P(A) = \int P(AB)\, dB$

Posterior: *Compute the posterior:*

$$p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta)\, p(\theta)$$

$\int P(\theta \mid x_{1:n})\, d\theta = 1$

$$P(\theta \mid x_{1:n}) = \frac{P(x_{1:n} \mid \theta)\, P(\theta)}{\int P(x_{1:n} \mid \theta)\, P(\theta)\, d\theta}$$

Marginal likelihood: $p(x_{1:n}) = \int P(x_{1:n} \mid \theta)\, P(\theta)\, d\theta$

Bayesian learning for coin model

Step 1: *Write down the likelihood of the data (i.i.d. Bernoulli in our case):*

$$p(x_i|\theta) = \Theta^{x_i}(1-\Theta)^{1-x_i} \qquad x_i \in \{0,1\}$$

$$p(x_{1:n}|\theta) = \Theta^{m}(1-\Theta)^{n-m} \qquad m = \# \text{ 1's}$$

Bayesian learning for coin model

Step 2: *Specify a prior on* θ*. For this, we need to introduce the Beta distribution.*

We know θ is continuous and $0 \leq \theta \leq 1$

$P(\theta)$?

$\theta = P(X_i = 1) \; \forall i$

$$P(\theta) = \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

(α, β) hyperparams

$$\Gamma(z) = \int_0^\infty e^x x^{z-1} dx$$

$$\log(z) = \int_1^z \frac{1}{x} dx$$

$$\int \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

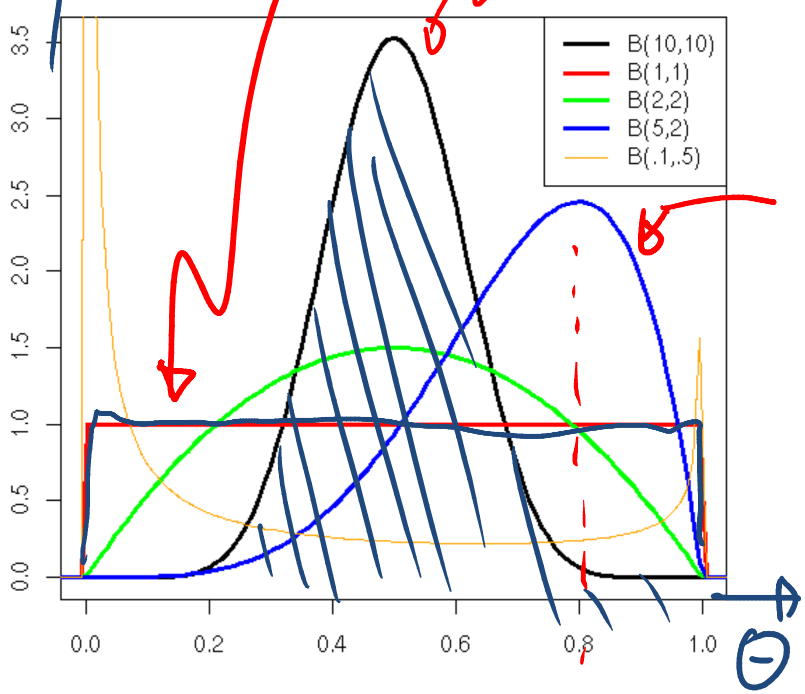
$$\int P(\theta)\, d\theta = 1 = \int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

Bayesian learning for coin model

Step 2: *Specify a prior on θ. For this, we need to introduce the Beta distribution.*

uniform

$P(\theta)$ $\quad$ $\theta \approx 0.5$



0.8

If $\alpha = 1 \quad \beta = 1$

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{0}(1-\theta)^{0}$$

$$= 1$$

i.e. $P(\theta) \propto 1$ (uniform)

$$\text{mean}(\theta) = \frac{\alpha}{\alpha+\beta}$$

Bayesian learning for coin model

Step 3: *Compute the posterior:*

$$p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta)\, p(\theta) =$$

$$= \theta^{m}(1-\theta)^{n-m}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}$$

$$= \theta^{\alpha'-1}(1-\theta)^{\beta'-1} \qquad \alpha' = m+\alpha$$

$$\beta' = n-m+\beta$$

$$p(\theta \mid x_{1:n}) = \frac{1}{const}\,\theta^{\alpha'-1}(1-\theta)^{\beta'-1}$$

$$= \frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}\,\theta^{\alpha'-1}(1-\theta)^{\beta'-1}$$

Example

Suppose we observe the data, $\mathbf{x_{1:6}} = \{\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}\}$*, where each* $\mathbf{x_i}$ *comes from the same Bernoulli distribution (i.e. it is independent and identically distributed (**iid**)). What is a good guess of* θ*?*

$\hat{\theta}_{ML} = 1$

We can compute the posterior and use its mean as the estimate.

$$P(\theta | x_{1:6}) \propto \theta^{6+\alpha-1} (1-\theta)^{0+\beta-1}$$

$$\hat{\theta}_B = \mathbb{E}(\theta | x_{1:6}) = \frac{6+\alpha}{6+\alpha+\beta}$$

1000000 1's

$\hat{\theta}_B \approx 1$

Using a prior Beta(2,2):

$$\hat{\theta}_B = \frac{8}{10} = 0.8$$

Beta(1,1)

$$\hat{\theta}_B = \frac{7}{7+1} = \frac{7}{8}$$

Using a prior Beta(1,0.01):

$$\hat{\theta}_B = \frac{7}{7+0.01} \approx 1$$

Next lecture

In the next lecture, we apply our learning strategies to Bayesian networks.