Bayesian Machine Learning

MAP vs Max Likelihood

Marek Petrik

3/2/2017

Bayesian Machine Learning

- Maximum likelihood
- What if we have prior knowledge?
- Improve on maximum likelihood

Estimating Coefficients: Maximum Likelihood

- **Likelihood**: Probability that data is generated from a model

$$\ell(\text{model}) = \Pr[\text{data} \mid \text{model}]$$

- Find the most likely model:

$$\max_{\text{model}} \ell(\text{model}) = \max_{\text{model}} \Pr[\text{data} \mid \text{model}]$$

- Likelihood function is difficult to maximize
- Transform it using $\log$ (strictly increasing)

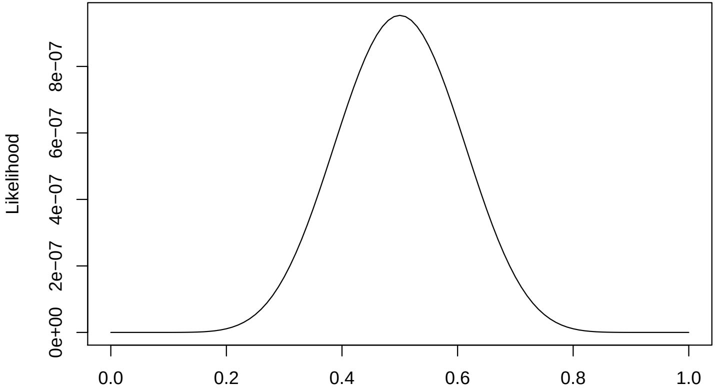
$$\max_{\text{model}} \log \ell(\text{model})$$

- Strictly increasing transformation does not change maximum

Example: Maximum Likelihood

- Assume a coin with p as the probability of *heads*
- **Data**: h heads, t tails
- The likelihood function is:

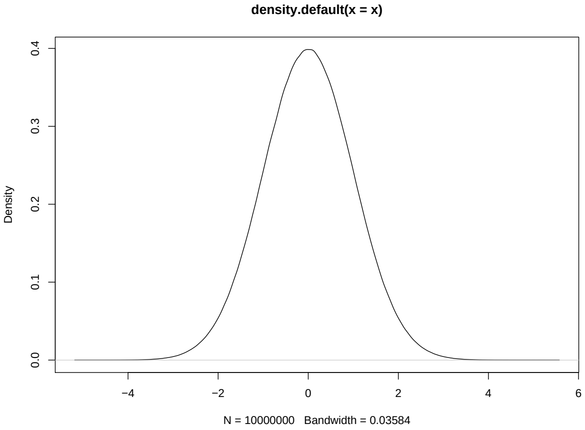
$$\ell(p) = \binom{h+t}{h} p^h (1-p)^t \approx p^h (1-p)^t .$$



Normal Distribution

- Normal density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Maximum Likelihood For OLS

- Assume $\epsilon_i \sim \mathcal{N}(0, 1)$
- Likelihood of a single data point

$$f(y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - \hat{y}_i)^2}{2}}$$

- Recall

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- Likelihood of all data

$$\prod_{i=1}^{n} f(y_i)$$

Problems with Maximum Likelihood

- Example!

Bayes Theorem

- Classification from label distributions:

$$\Pr[Y = k \mid X = x] = \frac{\Pr[X = x \mid Y = k] \Pr[Y = k]}{\Pr[X = x]}$$

- Example:

$$\Pr[\text{default} = \text{yes} \mid \text{balance} = \$100] = \frac{\Pr[\text{balance} = \$100 \mid \text{default} = \text{yes}] \Pr[\text{default} = \text{yes}]}{\Pr[\text{balance} = \$100]}$$

- Notation:

$$\Pr[Y = k \mid X = x] = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Better Options

1. **Maximum likelihood**

$$\max_{\text{model}} \Pr[\text{data} \mid \text{model}]$$

2. **Maximum a posteriori estimate (MAP)**

$$\max_{\text{model}} \Pr[\text{model} \mid \text{data}]$$

Better Options

1. **Maximum likelihood**

$$\max_{\text{model}} \Pr[\text{data} \mid \text{model}]$$

2. **Maximum a posteriori estimate (MAP)**

$$\max_{\text{model}} \Pr[\text{model} \mid \text{data}] = \max_{\text{model}} \Pr[\text{data} \mid \text{model}] \frac{\Pr[\text{model}]}{\Pr[\text{data}]}$$

Better Options

1. **Maximum likelihood**

$$\max_{\text{model}} \Pr[\text{data} \mid \text{model}]$$

2. **Maximum a posteriori estimate (MAP)**

$$\max_{\text{model}} \Pr[\text{model} \mid \text{data}] = \max_{\text{model}} \Pr[\text{data} \mid \text{model}] \frac{\Pr[\text{model}]}{\Pr[\text{data}]}$$

Identical when the prior is normal

Maximum a Posteriori Estimate

$$\Pr[\beta \mid X, Y] = \alpha f(Y \mid X, \beta) p(\beta \mid X) = \alpha f(Y \mid X, \beta) p(\beta)$$

- *Prior*: $p(\beta)$
- *Likelihood*: $f(Y \mid X, \beta)$

Better Solution

- Example!