# Example of MSE/MAP Estimation

Note Title 1/18/2013

Consider a Bernoulli distribution

$$f(x_i|u) = u^{x_i}(1-u)^{1-x_i} \qquad 0 \le u \le 1$$

This describes a distribution over binary values 0, 1 that could represent Heads/Tails, yes/no, or any other two-state event.

We will rewrite this likelihood function in a more general way, to make it easier to generalize later to distributions over arbitrary # of discrete states.

Let $u = u_1$ and $1-u = u_2$, so $u_1 + u_2 = 1$

Let $z_{i1} = \begin{cases} 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 0 \end{cases}$ $\quad z_{i2} = \begin{cases} 1 & \text{if } x_i = 0 \\ 0 & \text{if } x_i = 1 \end{cases}$

These $z_{ij}$ are binary indicator variables used to form a "1 of K" representation. [Chris Bishop's PRML book discusses this representation in more detail, if you have access to the book]

example $x = \{\overset{x_1}{H}, \overset{x_2}{T}, \overset{x_3}{T}, \overset{x_4}{H}, \overset{x_5}{H}\}$

| $z_{ik}$ | $k=1$ | $k=2$ |
|---|---|---|
| $i=1$ | 1 | 0 |
| $i=2$ | 0 | 1 |
| $i=3$ | 0 | 1 |
| $i=4$ | 1 | 0 |
| $i=5$ | 1 | 0 |
| | 3 | 2 |
| | $\sum z_{i1} = N_1$ | $\sum z_{i2} = N_2$ |

Note for each row, only one variable is 1 and the other is 0. In a general 1 of K situation, one variable is 1 and the other K−1 variables are 0.

Furthermore, note sum down each column is # samples taking value k using notation from previous page, we can now write our bernoulli distribution in more general form as:

$$P(x_i|u) = u_1^{z_{i1}} \; u_2^{z_{i2}}$$

Now, given N samples $X = \{x_1, x_2, \dots, x_N\}$, form the joint likelihood function for parameters $u = \{u_1, u_2\}$

$$L(u) = P(X|u) = \prod_{i=1}^{N} \prod_{k=1}^{2} u_k^{z_{ik}} = \prod_{i=1}^{N} u_1^{z_{i1}} u_2^{z_{i2}}$$

Now solve for MLE estimate

Lagrange multiplier to enforce constraint $u_1 + u_2 = 1$

$$\log L = \sum_{i=1}^{N} z_{i1} \log u_1 + z_{i2} \log u_2 + \lambda(1 - u_1 - u_2)$$

$$\frac{\partial \log L}{\partial u_1} = \sum_{i=1}^{N} \frac{z_{i1}}{u_1} - \lambda = 0 \implies \sum z_{i1} = N_1 = \lambda u_1$$

$$\frac{\partial \log L}{\partial u_2} = \sum_{i=1}^{N} \frac{z_{i2}}{u_2} - \lambda = 0 \implies \sum z_{i2} = N_2 = \lambda u_2$$

$$\frac{\partial \log L}{\partial \lambda} = 1 - u_1 - u_2 = 0 \implies u_1 + u_2 = 1$$

sum both sides

$$N_1 + N_2 = \lambda \underbrace{(u_1 + u_2)}_{1}$$

so
$$u_1 = N_1 / \lambda = N_1 / (N_1 + N_2)$$
$$u_2 = N_2 / \lambda = N_2 / (N_1 + N_2)$$

NOTE: The MLE estimates are just the relative frequency of counts of values occurring in the sample data.

To compute MAP estimates, recall Bayes rule

$$\underbrace{P(u|x)}_{\text{posterior}} = \frac{\overbrace{P(x|u)}^{\text{likelihood}}\ \overbrace{P(u)}^{\text{prior}}}{\underbrace{P(x)}_{\text{evidence}}}$$

Since the denominator $P(x)$ does not depend on $u$, we can ignore it if we only care about $\arg\max P(u|x)$.

so $\hat{u}_{MAP} = \arg\max_u P(u|x) = \arg\max_u P(x|u)P(u)$

From MLE derivation on previous page, we know that the joint likelihood $P(x|u)$ is

$$P(x|u) = \prod_{i=1}^{N} u_1^{z_{i1}} u_2^{z_{i2}} = u_1^{\Sigma z_{i1}} u_2^{\Sigma z_{i2}} = u_1^{N_1} u_2^{N_2}$$

with $u_1 + u_2 = 1$, [note: our previous MLE derivation could have been streamlined by writing the joint likelihood function in this way.]

We would like to multiply this by a "conjugate prior" that does not greatly complicate the form of the product. The beta distribution is a conjugate prior for the bernoulli distribution.

$$\mathrm{Beta}(u|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \underbrace{u}_{\text{our } u_1}{}^{a-1} \underbrace{(1-u)}_{\text{our } u_2}{}^{b-1}$$

So let our prior be

$$P(u) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{} u_1^{a-1} u_2^{b-1}$$

normalizing constant. Does not depend on $u$ parameters, so we will drop it from now on.

Now form the posterior

$$p(u|x) \propto p(x|u)\,p(u) = u_1^{N_1} u_2^{N_2} u_1^{a-1} u_2^{b-1}$$

$$= u_1^{N_1+a-1} \quad u_2^{N_2+b-1}$$

Now want $\underset{u_1,u_2}{\text{argmax}}\; p(u|x)$, subject to $u_1+u_2=1$

$$\log p(u|x) = (N_1+a-1)\log u_1 + (N_2+b-1)\log u_2 + \lambda(1-u_1-u_2)$$

$$\frac{\partial \log p}{\partial u_1} = \frac{(N_1+a-1)}{u_1} - \lambda = 0$$

$$\frac{\partial \log p}{\partial u_2} = \frac{N_2+b-1}{u_2} - \lambda = 0$$

Note! in the homework you will fill in the details, for a more general distribution

$$\frac{\partial \log p}{\partial \lambda} = 1-u_1-u_2=0$$

$$\Longrightarrow u_1+u_2=1$$

Solving for $u_1$ and $u_2$ in a similar manner to our MLE derivation, we find

$$u_1 = \frac{N_1+a-1}{N_1+N_2+a+b-2} \qquad u_2 = \frac{N_2+b-1}{N_1+N_2+a+b-2}$$

Intuitive interpretation: $a-1$ and $b-1$ are "virtual" sample counts given to us prior to observing the actual data, so we pool together both the real and virtual samples.