Week 7: Bayesian Inference and Parameter Estimation

Phong Le, Willem Zuidema

December 11, 2013

We are curious about some events or things (such as a language) and want to study their *hidden* mechanisms (grammar) G_{true}. A proper way to do is to collect a lot of data (sentences, dialogues) $D = \{x_1, x_2, ..., x_n\}$ and then find a model $\hat{G}$ that best *fits* (or explains) D. In this way, you expect that $\hat{G}$ is a 'good' estimate of G_{true}.

In this lab, firstly, we will study one quality metric to measure the 'degree of belief' that a model G is a good estimate of G_{true} given observed data D: the posterior probability $P(G|D)$, and how to compute it by using Bayesian inference. Then, we will examine two widely used estimation methods: Maximum Likelihood estimation (MLE) and Maximum A Posteriori estimation (MAP).

Required R Code At `http://www.illc.uva.nl/LaCo/clas/fncm13/assignments/computerlab-week7/` you can find the R-files you need for this exercise.

1 Bayesian Inference

In statistics, according to Wikipedia, Bayesian inference

> is a method of inference in which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is acquired.

In other words, Bayesian inference is to compute the posterior probability $P(G|D)$ based on the Bayes' rule

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \tag{1}$$

where $P(G)$ is the prior probability of G and D is additional evidence. In order to illustrate the method, let's examine the toy example below.

Toy Example: Murder in Dam Square

A man was found dead in Dam Square and two people, namely A and B, are suspected. After 24h investigating, the police found four witnesses, one of them reported that he saw A shooting the victim whereas the others said B. However, because it was foggy at that time, the police estimate that those witnesses only 80% correctly distinguished the two suspects. Our task is using Bayesian inference to help the police find out which one is the murderer, A or B.

First of all, we need to model the problem mathematically. Let's denote

- $P(X)$ the prior probability that X is the murderer (note: $P(X = B) = 1 - P(X = A)$)
- $P(W_i|X)$ ($i = 1..4$) the confidence of the i-th witness' vision. Here, $P(W_i = X|X) = 0.8$.
- $P(X|W_{1,2,3,4})$ the posterior probability that X is the murderer based on the evidence given by all the four witnesses.

Our goal is to compute the posterior probability $P(X = A|W_1 = A, W_2 = B, W_3 = B, W_4 = B)$ by updating the posterior probability when additional evidence is given as follows

- Step 0: when we don't have any evidence, we can only judge based on the prior probability $P(X)$.
- Step 1: after the first witness reports, we update the posterior probability

$$P(X|W_1 = A) = \frac{P(W_1 = A|X)P(X)}{P(W_1 = A)}$$

where $P(W_1 = A) = \sum_{X\in\{A,B\}} P(W_1 = A|X)P(X)$.

Exercise 1.1: We set up the experiment as follows

```
p.prior = c(0.5,0.5)                          # P(X=A) = P(X=B) = 0.5
likelihood = matrix(c(0.8,0.2,0.2,0.8),2,2)   # P(W_i = X | X) = 0.8
witness = c(1,2,2,2)                          # W1 = A, W2 = W3 = W4 = B
```

where we represent the likelihood-function as a matrix that gives for each actual killer (A,B) the likelihood of obtaining a witness-report incriminating A or B. Calculate (in R) the probability that A or B is the killer before and after hearing witness 1.

we then continue with incorporating the information from witnesses 2, 3 and 4. Note that the posterior after witness 1 becomes the prior for caclulating the posterior after witness 2!

- Step 2: after the second witness reports, we update the posterior probability

$$P(X|W_1 = A, W_2 = B) = \frac{P(W_2 = B|X)P(X|W_1 = A)}{P(W_2 = B|W_1 = A)}$$

where $P(X|W_1 = A)$ is computed in step 1. (Note: because W_i, W_j with $i \neq j$ are independent given X, $P(W_2 = B|X, W_1) = P(W_2 = B|X)$.)

- Step 3: after the third witness reports, we update the posterior probability

$$P(X|W_1 = A, W_2 = B, W_3 = B) = \frac{P(W_3 = B|X)P(X|W_1 = A, W_2 = B)}{P(W_3 = B|W_1 = A, W_2 = B)}$$

where $P(X|W_1 = A, W_2 = B)$ is computed in step 2.

- Step 4: after the last witness reports, we update the posterior probability

$$P(X|W_1 = A, W_2 = B, W_3 = B, W_4 = B) = \frac{P(W_4 = B|X)P(X|W_1 = A, W_2 = B, W_3 = B)}{P(W_4 = B|W_1 = A, W_2 = B, W_3 = B)}$$

where $P(X|W_1 = A, W_2 = B, W_3 = B)$ is computed in step 3.

Exercise 1.2. The script `murder.R` automatizes the calculations at step 0-4.

- Step 0:

```
  # step 0
  p.poste = p.prior
  print(p.poste)
```

- Step 1, 2, 3, 4:

```
  # step i > 0
  for (i in 1:length(witness)) {
      if (witness[i] == 1) # if the witness saw A
          p.poste = p.poste * c(p.witness,1-p.witness)
```

```
        else # if the witness saw B
            p.poste = p.poste * c(1-p.witness,p.witness)
        p.poste = p.poste / sum(p.poste)  # normalize

        print(p.poste)
}
```

Is the posterior probability at step 2 the same step 0? Explain why?
Based on the posterior probability after step 4, who is the most suspected?

Exercise 1.3: In exercise 1, the prior distribution is uniform, because we haven't had any evidence yet. Now, assuming that B is a law-abiding citizen according to all records, whereas A has prior convictions for violence and other crimes. It might therefore be reasonable to suspect A more than B. We adjust the prior distribution as follows

```
p.prior = c(0.9,0.1)   # P(X=A) = 0.9, P(X=B) = 0.1
```

while keeping other parameters unchanged. Compute the posterior distribution as in exercise 1 and report what you get.

2 Parameter Estimation

In the previous section, we study how to use Bayesian inference to estimate a distribution. In this section, we will study how to select the 'best' model given observed data.

Maximum Likelihood Estimation (MLE) is a method to find values for model's parameters such that the likelihood given the observed data, e.g. the probability of the observed data given the model, is maximized

$$\hat{G}_{MLE} = \max_G P(D|G) \tag{2}$$

Maximum A Posteriori (MAP) Estimation on the other hand, is to maximize the posterior probability

$$\hat{G}_{MAP} = \max_G P(G|D) \tag{3}$$

According to the Bayes' theorem, we can compute posterior probability based on prior probability and likelihood, e.g. $P(G|D) = \frac{P(D|G)P(G)}{P(D)}$. Therefore

$$\hat{G}_{MAP} = \max_G \frac{P(D|G)P(G)}{P(D)} = \max_G P(D|G)P(G) \tag{4}$$

(because $P(D)$ is a constant in this case, we freely drop it).

In order to easily compute $P(D|G)$ in Equation 2 and 4, observed data are assumed to be *independent and identically distributed* (i.i.d), e.g. examples are independently drawn from the same distribution. Hence

$$P(D = \{x_1, x_2, ..., x_n\}|G) = \prod_{i=1}^{n} P(x_i|G) \tag{5}$$

Exercise 2.1. What are the MLE and MAP hypotheses in exercise 1.3 after 4 witness reports? And what were they after the first 3 witness reports?

Because probabilities can become very small and multiplication is a relatively expensive operation, it is often convenient to work with the logarithm of probabilities.

Exercise 2.2. Confirm in R that :

$$\prod_i p_i = \exp \sum_i \log p_i$$

Now, Equation 2 and 4 respectively become [1]

$$\hat{G}_{MLE} = \max_G \prod_{i=1}^{n} P(x_i|G) = \max_G \sum_{i=1}^{n} \log P(x_i|G) \tag{6}$$

where the right hand side, $\sum_{i=1}^{n} \log P(x_i|G)$, is called *log-likelihood*, and

$$\hat{G}_{MAP} = \max_G P(G) \prod_{i=1}^{n} P(x_i|G) = \max_G \big(\log P(G) + \sum_{i=1}^{n} \log P(x_i|G) \big) \tag{7}$$

Toy Example

In the following exercises, we will examine a very simple case: estimating the mean of a normal distribution $N(x; \mu, \sigma^2)$. The scenario is that, we draw a sample $D = \{x_1, ..., x_n\}$ from $N(x; \mu_{true}, \sigma^2_{true})$; then, we ask you to estimate μ_{true}. (Note that, in order to adapt the above equations, we need to replace probability by density.)

Note that, by the definition of a normal distribution, if x is distributed according to a normal distribution with mean μ and standard deviation σ (i.e., $x \sim N(\mu, \sigma^2)$) then

$$p(x|\mu) = \frac{1}{2\sigma\sqrt{\pi}} \exp\big(-\frac{(x-\mu)^2}{2\sigma^2} \big) \tag{8}$$

which can be rewritten as

$$\log p(x|\mu) = -\frac{(x-\mu)^2}{2\sigma^2} + U \tag{9}$$

where U is a constant independent from μ (and can often be, conveniently, ignored). Now, Equation 6 and 7 respectively become

$$\hat{\mu}_{MLE} = \max_\mu - \sum_{i=1}^{n} (x_i - \mu)^2 \tag{10}$$

$$\hat{\mu}_{MAP} = \max_\mu \big(\log p(\mu) - \sum_{i=1}^{n} (x_i - \mu)^2 \big) \tag{11}$$

Exercise 2.3: The file 'estimate_mu.R' provides you with a visualization tool for the estimation problem (with both MLE and MAP): each time you press the Enter key, the program will draw an example from the true model and use it to update $\hat{\mu}_{MLE}$ and $\hat{\mu}_{MAP}$; after that, it will show a plot containing graphs of log-likelihood and log posterior probability over μ and another plot containing graphs of $\hat{\mu}_{MLE}$ and $\hat{\mu}_{MAP}$ over sample size.

In this exercise, we assume that the prior distribution is also a normal distribution $p(\mu) = N(\mu; \mu_\mu, \sigma^2_\mu)$

1. First of all, you need to set values for parameters and draw a sample

[1] Note that because log is a monotonically increasing function, $\max(a, b) = \max\big(\log(a), \log(b)\big)$.

```r
mu.true = 3         # mean
sigma.true = 10     # standard deviation
n = 100             # sample size
data = rnorm(n, mean = mu.true, sd = sigmoid.true)

mean.mu = 2.5       # mean of mu (priori)
sd.mu = 1           # standard deviation of mu (priori)
```

2. Before executing the file, try to predict how the graph of log-likelihood over μ looks like, and how the graph of log-posterior-probability over μ looks like when (i) observed data are ignored and (ii) observed data are used.

3. Load the file (`source("estimate_mu.R")`), and then execute `estimate.mu(data, sigma.true, mean.mu, sd.mu, plot=T)` (note: the black lines are of MLE, the blue lines MAP). Report what you get.

4. It can be shown that $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Confirm that by computing the sample average `sum(data)/n`. (Note: $\hat{\mu}_{MLE}$ computed by the program is rounded.)

5. Change the prior $p(\mu)$ to have `mean.mu = -2, sd.mu = 1` then execute `estimate.mu(...)` again. Now set `mean.mu = -2, sd.mu = 1000` then execute `estimate.mu(...)`. Do you have any conclusion about the effect of the prior distribution?

Exercise 2.4 (optional): In this exercise, we will compare MLE to MAP by computing mean squared errors over sample size.

1. First, we set up the experiment as in exercise 1

```r
n = 100
mu.true = 3
sigma.true = 10
mean.mu = 2.5
sd.mu = 1
```

Then, we compute mean squared errors of m runs

```r
mse.mle = rep(0,n); mse.map = rep(0,n)
m = 100

for (i in 1:m) {
    data = rnorm(n, mean = mu.true, sd = sigma.true)
    mu.est = estimate.mu(data, sigma.true, mean.mu, sd.mu, plot=F)
    mse.mle = mse.mle + (mu.est$mu.mle - mu.true)^2
    mse.map = mse.map + (mu.est$mu.map - mu.true)^2
}

mse.mle = sqrt(mse.mle) / m
mse.map = sqrt(mse.map)  / m
```

And finally plot the errors

```r
plot(1:n, mse.mle, type='l',ylim=c(min(min(mse.mle),min(mse.map)),max(
     max(mse.mle),max(mse.map))), xlab = 'sample size', ylab = 'MSE')
lines(1:n, mse.map, col='blue')
```

(Don't forget our notation: black is of MLE and blue MAP.)

2. Set `n = 3000` and rerun the above.

3. Based on what you have done so far, draw conclusions about MLE vs MAP and when MAP is useful.

3 Submission

You have to submit a file named 'your_name.pdf'. The deadline is 15:00 Monday 16 Dec. If you have any questions, contact Phong Le (p.le@uva.nl).