Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων

ΔΠΜΣ στις Εφαρμοσμενες Μαθηματικες Επιστημες

Μπεϋζιανή Στατιστική και MCMC

Μάιος 30, 2017

Φοιτητής:
Παντελής Καρατζας
email: pantelispanka@gmail.com

Διδάσκων:
ΔΗΜΗΤΡΗΣ ΦΟΥΣΚΑΚΗΣ
email: fouskakis@math.ntua.gr

Άσκηση 1

Τα δεδομένα της άσκησης μας δίνουν την διάρκεια ζωής (σε ώρες) 20 ηλεκτρονικών εξαρτημάτων.
Οι τιμές είναι οι παρακάτω:

$$60, 119, 100, 130, 43, 227, 23, 91, 128, 199, 85, 125, 40, 26, 141, 212, 238, 94, 111, 67$$

Αναλυτικός υπολογισμός ύστερης κατανομής

Για να υπολογίσουμε αναλυτικά την *posterior* κατανομή χρειαζόμαστε την συνάρτηση πιθανοφάνειας της κατανομής που ακολουθούν τα δεδομένα, (εδώ X $\sim N(\mu, \sigma^2)$ με άγνωστη μέση τιμή και τυπική απόκλιση $\sigma = 60$), καθώς και την prior. Γνωρίζουμε ότι η πρότερη πληροφορία μας δίνεται πάλι από μια Κανονική κατανομή με μέση τιμή $\mu = 0$ και διασπορά $\sigma^2 = 1000$, ”περικομένη” στο διάστημα $(0, 500)$. Εφόσον και οι δυο κατανομές είναι της ίδιας οικογένειας (εκθετικές κατανομές) και άρα έχουμε μία $conjugateprior$ κατανομή το αποτέλεσμα θα είναι μια κανονική κατανομή πάλι. Μένει να υπολογίσουμε τις μ και σ της *posterior* κατανομής.

$$\text{Δεδομένα}: \quad N(\mu, \sigma^2) : f(x_n \mid \mu.\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ - \frac{(x-\mu)^2}{2\sigma^2} \Big\} \quad n = 1, 2, ..., 20$$

$$Likelihood: \quad N(\mu, \sigma^2) : L(X \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ - \frac{(x-\mu)^2}{2\sigma^2} \Big\} \quad using \qquad \tau = \frac{1}{\sigma^2}$$

$$= \big(\frac{\tau}{2\pi}\big)^{\frac{n}{2}} exp\Big\{ - \frac{1}{2}\tau \sum_1^n (x_n - \mu)^2 \Big\} = \big(\frac{\tau}{2\pi}\big)^{\frac{n}{2}} exp\Big\{ - \frac{1}{2}\tau \Big(\sum_1^n (x_n - \bar{x})^2 - n(\bar{x} - \mu)^2 \Big) \Big\}$$

$$Prior: \qquad f(\mu) = \begin{cases} \frac{1}{\sigma_0\sqrt{2\pi}\left[\Phi((\beta-\mu_0)/\sigma_0) - \Phi((\alpha-\mu_0)/\sigma_0)\right]} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} & \text{if } \alpha \le \mu \le \beta \\ 0 & \text{else.} \end{cases}$$

$$Posterior: \quad p(\mu \mid \mathrm{X}) \propto p(X \mid \mu) p(\mu) \propto const * exp\Big\{ - \frac{1}{2}\Big(\tau\Big(\sum_1^n (x_n - \bar{x})^2 - n(\bar{x} - \mu)^2 \Big) + \tau_0 \big(\mu - \mu_0\big)^2 \Big) \Big\}$$

$$\propto exp\Big\{ - \frac{1}{2}\Big(n\tau(\bar{x} - \mu)^2 + \tau_0\big(\mu - \mu_0\big)^2 \Big) \Big\}$$

$$= exp\Big\{ - \frac{1}{2}(n\tau + \tau_0)\Big(\mu - \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}\Big)^2 + \frac{n\tau\tau_0}{n\tau+\tau_0}(\bar{x} - \mu_0) \Big\}$$

$$\propto exp\Big\{ - \frac{1}{2}(n\tau + \tau_0)\Big(\mu - \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}\Big)^2 \Big\} \text{ if } \alpha \le \mu \le \beta$$

Από τον αναλυτικό υπολογισμό της ύστερης κατανομής έχουμε : $p(\mu \mid X) \sim N(\frac{n\tau\bar{x}+\tau_0\mu_0}{n\tau+\tau_0}, \frac{1}{n\tau+\tau_0})$. Ο μέσος των δεδομένων μας ακολουθεί κανονική κατανομή με παραμέτρους όπως έχουν υπολογιστεί. Για τον αναλυτικό υπολογισμό και αυτών έχουμε $\tau = \frac{1}{\sigma^2}$, άρα $t_0 = 0.001$ και $t = 0.0002777$. Με αντικατάσταση των τιμώς στους τύπους του μέσου και διασποράς πέρνουμε τις τιμές $\mu = 95.57308$ και $\sigma^2 = 153.8462$

Διαγράμματα Πιθανοφάνειας, πρότερης και ύστερης κατανομής

Παρακάτω εμφανίζεται ένα συγκεντρωτικό διάγραμμα.
Σε αυτό βλέπουμε το διάγραμμα των:

· **Διάγραμμα πιθανοφάνειας** (χρησιμοποιείται το κόκκινο χρώμα)

· **Διάγραμμα** *Prior* **κατανομής** (χρησιμοποιείται το καφέ χρώμα)

· **Διάγραμμα** *Posterior* **κατανομής** (χρησιμοποιείται το μαύρο χρώμα)

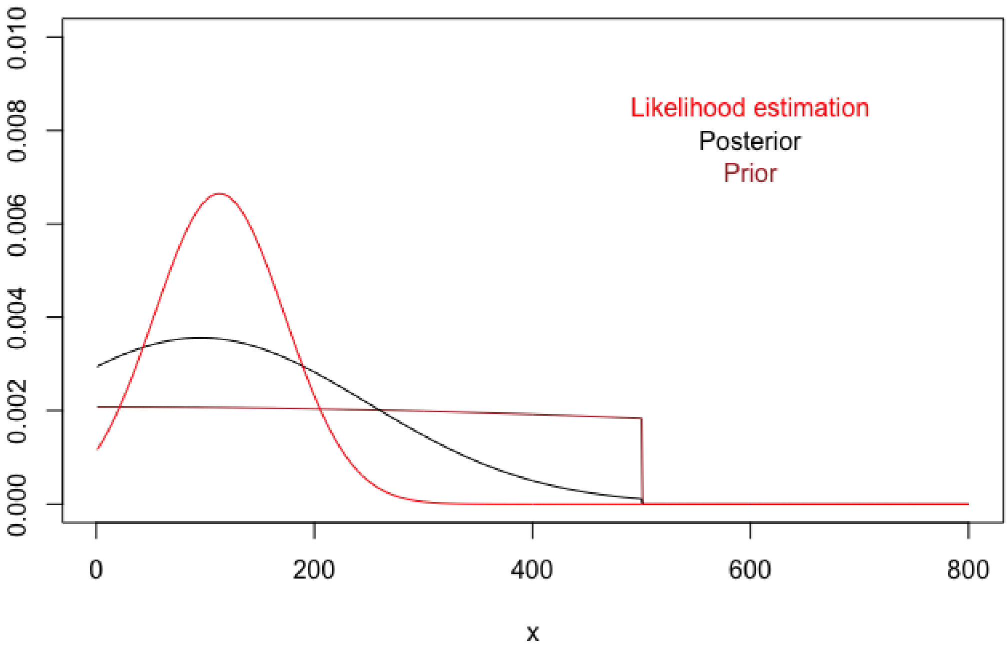

Figure 1: Διάγραμμα Πιθανοφάνειας, πρότερης και ύστερης κατανομής

Η σύγκριση που πρέπει να γίνει εδώ είναι αυτή των διαγραμμάτων εκτίμησης του μέσου με την μέθοδο της μέγιστης πιθανοφάνειας χωρίς την εισαγωγή "πρότερης" γνώσης στην διαδικασία μας και στο διάγραμμα της *posterior* κατανομής που δημιουργήθηκε με την εισαγωγή πρότερης γνώσης. Το τρίτο διάγραμμα της *prior* κατανομής αντιπροσωπεύει αυτό το κομμάτι που ως παλαιότερη πληροφορία είναι γνωστό και χρησιμοποιούμε για να φτειάξουμε την *posterior* κατανομή. Εύκολα μπορεί κανείς να διαπιστώσει το πώς αλλά και πόσο αυτή η πληροφορία επιδρά στον τρόπο υπολογισμού ενός μοντέλου για την μέση τιμή της διάρκειας ζωής των ηλεκτρονικών εξαρτημάτων που τώρα μελετάμε.

Η μέση τιμή χωρίς να χρησιμοποιήσουμε την *prior* κατανομή μεγαλώνει ελάχιστα και η διασπορά μικραίνει με αποτέλεσμα οι τιμές να είναι κατά βάση γύρω από την μέση τιμή σε αντίθεση με το μοντέλο που χρησιμοποιεί την πρότερη κατανομή. Σε αυτό (posterior) βλέπουμε πως υπάρχει πιθανότητα να υπάρξει εξάρτημα που η μέση τιμή της ζωής του να υπερβεί τις 250 ώρες σε αντίθεση με το προηγούμενο μοντέλο που αυτή σχεδόν μηδενίζεται. Η ασάφεια που προηγείται μέσο της μεγάλης διασποράς είναι αυτη που κυρίαρχει στο νέο μοντέλο (*posterior*) και φαίνεται πως είναι σημαντική η επίδρασή του σε αυτό, αλλοιώνοντας και τα δύο μεγέθη, της διασποράς αλλά και της μέσης τιμής.

Κώδικας που χρησιμοποιήθηκε

Για να δημιουργηθούν τα παραπάνω διαγράμματα και να υπολογιστούν οι απαραίτητες τιμές χρησιμοποιήσαμε τον παρακάτω κώδικα στην γλώσσα προγραμματισμού R.

```
#The Data of the problem
data <- c(60, 119, 100, 130, 43, 227, 23,
         91, 128, 199, 85, 125, 40, 26, 141, 212, 238, 94, 111, 67)



#Creating x values
x <- seq(1,800,by=1)

#Plotting Prior Distribution (truncated Normal N(0,1000) in 0-500)
plot(x, dtnorm(x,0,1000, 0 , 500), type = "l", ylim = c(0,0.004))

#Adding the Posterior with mean = 95.573, variance = 153.8462 As calculated above
lines(x, dtnorm(x,95.573,153.8462, 0 , 500), type = "l")

#Likelihood function of normal dist with known variance
likelihood.normal.mu <- function(mu, sig2, x) {
     # mu mean of normal distribution for given sig
     # sig2 variance of normall distribution
     # x vector of data
     n = length(x)
     a1 = (2*pi*sig2)^-(n/2)
     a2 = -1/(2*sig2)
     y = (x-mu)^2
     ans = a1*exp(a2*sum(y))
     return(ans)
 }

#Intitialize new empty array for the results of the likelihood
likelihood_results <- c()

#Gathering results from the likelihood estimator
for (i in 1:length(x)){
  likelihood_results[i] <- likelihood.normal.mu(mean(data), sig2=3600, x = x[i])
}

#Plotting Likelihood , adding Posterior, Prior
plot(x, dtnorm(x,0,1000, 0 , 500), type = "l", ylim = c(0,0.01), col = "brown")
lines(x, dtnorm(x,95.573,153.8462, 0 , 500), type = "l")
lines(x, likelihood_results, type = "l", col = "red")

#Adding text for better visualization
text(x=600,y=0.0085, label = "Likelihood estimation", col = "red")
text(x=600,y=0.0078, label = "Posterior")
text(x=600,y=0.0071, label = "Prior", col = "brown")
```

Metropolis Hastings Algorithm