Κεφάλαιο 23

Παραδείγματα Μεθόδων Ε-Μ Αλγόριθμου

Οι μέθοδοι Ε-Μ αλγόριθμου μπορούν να επεξηγηθούν πιο εύκολα στην περίπτωση ενός τυχαίου δείγματος το οποίο αποτελείται από παρατηρηθείσες και μη παρατηρηθείσες ή εκλειπούσες τιμές.

Ένα απλό παράδειγμα δείγματος με εκλειπούσες τιμές προκύπτει στην περίπτωση ελέγχου του χρόνου επιβίωσης. Για παράδειγμα, ένας αριθμός ηλεκτρικών λαμπτήρων ανάβει συνεχώς και καταμετρείται ο χρόνος που χρειάζεται μέχρι να πάψουν να λειτουργούν. Σε ένα τέτοιο παράδειγμα, είναι συνήθες φαινόμενο το πείραμα να διακοπεί πριν να πάψουν να λειτουργούν όλοι οι λαμπτήρες. Ο χρόνος επιβίωσης των λαμπτήρων οι οποίοι συνεχίζουν να δουλεύουν δεν έχει παρατηρηθεί. Όμως, προφανώς ο αριθμός των λογοκριμένων παρατηρήσεων και ο χρόνος της λογοκρισίας περιέχουν πληροφορία για την κατανομή του χρόνου επιβίωσης.

Ακόμη ένα γνωστό παράδειγμα στο οποίο μπορεί να χρησιμοποιηθεί ο Ε-Μ αλγόριθμος είναι το πεπερασμένο μοντέλο μίξης κατανομών. Κάθε παρατήρηση προέρχεται από μία άγνωστη παρατήρηση ενός υποτιθέμενου συνόλου κατανομών. Οι εκλείπουσες τιμές προσδιορίζουν την κατανομή. Οι παράμετροι των κατανομών πρόκειται να εκτιμηθούν. Ένα παράπλευρο κέρδος της μεθόδου είναι ότι εκτιμάται σε ποια κατηγορία ανήκουν τα δεδομένα.

Τα ελλειπή δεδομένα μπορούν να είναι εκλείπουσες παρατηρήσεις της ίδιας τυχαίας μεταβλητής η οποία παράγει το δείγμα που παρατηρήθηκε, όπως στην περίπτωση του παραδείγματος λογοκρισίας, ή μπορούν να προέρχονται από μία διαφορετική τυχαία μεταβλητή η οποία σχετίζεται με κάποιο τρόπο με την τυχαία

μεταβλητή που έχει παρατηρηθεί.

Πολλές εφαρμογές της μεθόδου του Ε-Μ αλγόριθμου περιλαμβάνουν προβλήματα με ελλειπή δεδομένα, αλλά αυτό δεν είναι αναγκαίο. Συχνά, ο Ε-Μ αλγόριθμος μπορεί να εφαρμοστεί βασισμένος σε μία τεχνητή ¨ελλειπής¨ τυχαία μεταβλητή για να συμπληρώσει το δεδομένα που παρατηρήθηκαν.

23.1 Πρώτο Παράδειγμα: Πολυωνυμική Κατανομή

Ένα από τα πιο απλά παραδείγματα της μεθόδου Ε-Μ αλγορίθμου δόθηκε από τους Dempster, Laird και Rubin (1977). Έστω η πολυωνυμική κατανομή με τέσσερα πιθανά αποτελέσματα, η οποία έχει συνάρτηση πυκνότητας πιθανότητας,

$$p(x_1, x_2, x_3, x_4) = \frac{n!}{x_1!x_2!x_3!x_4!}{\pi_1}^{x_1}{\pi_2}^{x_2}{\pi_3}^{x_3}{\pi_4}^{x_4}$$

με $n = x_1 + x_2 + x_3 + x_4$ και $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$. Έστω ότι όλες οι πιθανότητες συσχετίζονται με μία παράμετρο θ, ως εξής:

$$\pi_1 = \frac{1}{2} + \frac{1}{4}\theta$$

$$\pi_2 = \frac{1}{4} - \frac{1}{4}\theta$$

$$\pi_3 = \frac{1}{4} - \frac{1}{4}\theta$$

$$\pi_4 = \frac{1}{4}\theta$$

όπου $0 \leq \theta \leq 1$.

Δεδομένου μιας παρατήρησης (x_1, x_2, x_3, x_4), η λογαριθμική συνάρτηση πιθανοφάνειας δίνεται από

$$l(\theta) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta).$$

Σκοπός είναι να εκτιμηθεί η παράμετρος θ. Η παράγωγος δίνεται από

$$\frac{d}{d\theta} l(\theta) = \frac{x_1}{2 + \theta} - \frac{x_2 + x_3}{1 - \theta} + \frac{x_4}{\theta}.$$

και σε αυτό το απλό παράδειγμα, η εκτιμήτρια μεγίστης πιθανοφάνειας για το θ μπορεί να υπολογιστεί επιλύοντας μία απλή πολυωνυμική εξίσωση.

Για να χρησιμοποιηθεί ο Ε-Μ αλγόριθμος για αυτό το παράδειγμα,μπορεί κάποιος να υποθέσει την πολυωνυμική πέμπτης τάξης, η οποία παράγεται χωρίζοντας την πρώτη τάξη της αρχικής πολυωνυμικής σε δύο με αντίστοιχες πιθανότητες $1/2$ και $\theta/4$. Η αρχική μεταβλητή x_1 είναι τώρα το άθροισμα της u_1 και της u_2.

Κάτω από αυτόν το μετασχηματισμό, η Ε.Μ.Π. του θ θεωρώντας το άθροισμα u_2+x4 (ή x_2+x_3) να είναι μία πραγμάτωση της διωνυμικής με $n = u_2+x_4+x_2+x_3$ και $\pi = \theta$ (ή $1-\theta$). Ωστόσο, το u_2 (ή το u_1) δεν είναι γνωστό. Συνεχίζοντας σαν να είχαμε παρατήρηση από μία πολυωνυμική με πέντε πιθανά αποτελέσματα, με δύο ελλείποντα στοιχεία, η λογαριθμική πιθανοφάνεια δίνεται από,

$$l_c(\theta) = (u_2 + x_4)\log(\theta) + (x_2 + x_3)\log(1-\theta),$$

και η εκτιμήτρια μεγίστης πιθανοφάνειας για το θ είναι ίση με

$$\frac{u_2 + x_4}{u_2 + x_2 + x_3 + x_4}.$$

Το Ε-βήμα του επαναληπτικού Ε-Μ αλγόριθμου συμπληρώνει την ελλείπουσα τιμή με την αναμενόμενη τιμή δεδομένου της τρέχουσας τιμής της παραμέτρου, $\theta^{(k)}$ και της τιμής που έχει παρατηρηθεί. Αυτή είναι μία διωνυμική τυχαία μεταβλητή σαν μέρος της x_1. Έτσι, με $\theta = \theta^{(k)}$,

$$E_{\theta^{(k)}}(u_2) = \frac{\frac{1}{4}x_1\theta^{(k)}}{\frac{1}{2} + \frac{1}{4}\theta^{(k)}} = u_2^{(k)}.$$

Πρέπει τώρα να μεγιστοποιηθεί η $E_{\theta^{(k)}}(l_c(\theta))$. Επειδή η $l_c(\theta)$ είναι γραμμική έχουμε,

$$E(l_c(\theta)) = E(u_2 + x_4)\log(\theta) + E(x_2 + x_3)\log(1-\theta).$$

Το μέγιστο επιτυγχάνεται όταν

$$\theta^{(k+1)} = \frac{u_2^{(k)} + x_4}{u_2^{(k)} + x_2 + x_3 + x_4}.$$

Πιο κάτω βλέπουμε πως εφαρμόζεται ο Ε-Μ αλγόριθμος για το πιο πάνω παράδειγμα στην R, θέτοντας $\theta^{(0)} = 0.10$:

```
> theta=0.1
> p1=1/2+theta/4
> p2=1/4-theta/4
> p3=1/4-theta/4
> p4=theta/4
> x<-rmultinom(1,size=100,prob=c(p1,p2,p3,p4))
```

```
>
> thetainitial=0.60  ####initial value, that is \theta_{0}
> it  <- 0  ####iterative count
> del <- 1  ####iterative adjustment
> thetaold=thetainitial  ###assign the initial value to theta_{1}
> while(abs(del) > 0.000001 && (it <- it+1) < 20) ##Loop for 20 iterations and
+                                                  ##prespecified presicion
+ {
+ u2=(x[1]*thetaold)/(2+thetaold)        ##Expectation step
+ thetanew=(u2+x[4])/(sum(x)-x[1]+u2)    ##Maximization step
+ del=thetanew-thetaold                  ##Calculate the difference
+                                        ##between two iterations
+ thetaold=thetanew                      ##assign thetanew to thetaold for
+                                        ##the recursions
+ cat(it, thetanew, "\n")                ##List the iterations
+ }
1 0.2333333
2 0.1391061
3 0.1047372
4 0.09068726
5 0.08467897
6 0.08206028
7 0.08090947
8 0.08040192
9 0.0801777
10 0.08007859
11 0.08003476
12 0.08001537
13 0.0800068
14 0.080003
15 0.08000133
16 0.08000059
```

Όπως παρατηρούμε, για να πάρουμε ακρίβεια στο έκτο δεκαδικό στοιχείο στην εκτίμησή μας χρειάστηκαν δεκαέξι επαναλήψεις, και η εκτιμήτρια μεγίστης πιθανοφάνειας για το θ υπολογίστηκε να είναι ίση με `0.08000059`, ενώ η αληθινή τιμή του θ είναι ίση με 0.10. Ξαντρέχωντας τον αλγόριθμο για $\theta^{(0)} = 0.2$ παίρνουμε τα

πιο κάτω αποτελέσματα:

```
1 0.3402985
2 0.2634429
3 0.2330699
4 0.2197439
5 0.2136309
6 0.2107696
7 0.2094176
8 0.2087760
9 0.2084709
10 0.2083256
11 0.2082564
12 0.2082235
13 0.2082078
14 0.2082003
15 0.2081967
16 0.208195
17 0.2081942
```

Είναι φανερό ότι σε αυτήν την περίπτωση για να γίνει η σύγκλιση σε ακρίβεια στο έκτο δεκαδικό στοιχείο χρειάστηκαν δεκαεπτά επαναλήψεις, καταλήγοντας στην τιμή `0.2081942` ως εκτίμηση του θ.

23.2 Δεύτερο Παράδειγμα: Παραλλαγή του Πειράματος Ελέγχου Επιβίωσης Χρησιμοποιώντας Εκθετικό Μοντέλο

Έστω ότι ο χρόνος επιβίωσης ενός λαμπτήρα ακολουθεί την εκθετική κατανομή με μέση τιμή θ. Για να εκτιμηθεί το θ, καταμετρήθηκε ο χρόνος n λαμπτήρων από την ώρα που ανάβουν για πρώτη φορά μέχρι να πάψουν να λειτουργούν, $x_1, \ldots, x_n$. Σε ένα άλλο πείραμα, ελέγχθηκαν m λαμπτήρες, αλλά αυτήν την φορά δεν καταμετρήθηκαν ξεχωριστά ο χρόνος επιβίωσης του κάθε λαμπτήρα, αλλά ο αριθμός των λαμπτήρων, r, οι οποίοι έπαψαν να λειτουργούν σε μία χρονική στιγμή t.

Τα ελλειπή δεδομένα είναι οι χρόνοι επιβίωσης των λαμπτήρων στο δεύτερο

πείραμα, $u_1, \ldots, u_m$. Τότε,

$$l_c(\theta; x; u) = -n(\log(\theta) + \bar{x}/\theta) - \sum_{i=1}^{m} (\log(\theta) + u_i/\theta).$$

Η αναμενόμενη τιμή του χρόνου επιβίωσης ενός λαμπτήρα που δεν έχει ακόμη πάψει να λειτουργεί είναι ίση με

$$t + \theta,$$

ενώ, κάποιου που έχει πάψει να λειτουργεί είναι ίση με

$$\theta - \frac{te^{-t/\theta^{(k)}}}{1 - e^{-t/\theta^{(k)}}}.$$

Συνεπώς, χρησιμοποιώντας μία προσωρινή τιμή $\theta^{(k)}$, και το γεγονός ότι οι r από τους m λαμπτήρες δεν λειτουργούν, έχουμε την $E_{U|x,\theta^{(k)}}(l_c)$ να δίνεται στη μορφή

$$q^{(k)}(x, \theta) = -(n + m)\log(\theta) - \frac{1}{\theta}(n\bar{x} + (m - r)(t + \theta^{(k)}) + r(\theta^{(k)} - th^{(k)})),$$

όπου,

$$h^{(k)} = \frac{e^{-t/\theta^{(k)}}}{1 - e^{-t/\theta^{(k)}}}.$$

Το k-οστό Μ βήμα ορίζει τη μέγιστη τιμή συναρτήσει της μεταβλητής θ, η οποία, δεδομένου του $\theta^{(k)}$, παρατηρείται στο σημείο

$$\theta^{(k+1)} = \frac{1}{n + m}(n\bar{x} + (m - r)(t + \theta^{(k)}) + r(\theta^{(k)} - th^{(k)})).$$

Ξεκινώντας με μία θετική τιμή $\theta^{(0)}$, η πιο πάνω εξίσωση επαναλαμβάνεται μέχρι να επέλθει η σύγκλιση. Η αναμενόμενη τιμή $q^{(k)}$ δε χρειάζεται να υπολογίζεται κάθε φορά. Για να δούμε πως δουλεύει ο αλγόριθμος, παράγονται αρχικά μερικά τεχνητά δεδομένα με τη βοήθεια της R:

```
> # Generate data from an exponential with theta=2, and with the second
> # experiment truncated at t=3. Note that R uses a form of the
> # exponential in which the parameter is a multiplier; i.e., the R
> # parameter is 1/theta. Set the seed, so computations are reproducible.
>
> set.seed(4)
> n<-100
> m<-500
> theta<-2
```

```
> t<-3
> x<-rexp(n,1/theta)
> r<-min(which(sort(rexp(m,1/theta))>=3))-1
```

Συνεχίζοντας, εφαρμόζεται ο Ε-Μ αλγόριθμος στην R χρησιμοποιώντας ως αρχική τιμή το $\theta^{(0)} = 1$

```
> # We begin with theta=1.
> # (Note that theta.k is set to theta.kp1 at the beginning of the loop.)
> theta.k<-.01
> theta.kp1<-1
> # Do some preliminary computations
> n.xbar<-sum(x)
> # Then loop and test for convergence
> it  <- 0  ####iterative count
> del <- 1  ####iterative adjustment
> while(abs(del) > 0.000001 && (it <- it+1) < 20) ###Loop for 20 iterations
+                                                 ###and prespecified presicion
+ {
+ theta.kp1<-(n.xbar+
+ (m-r)*(t+theta.k)+
+ r*(theta.k-
+ t*exp(-t/theta.k)/(1-exp(-t/theta.k))
+   )
+ )/(n+m)
+ del<- theta.kp1-theta.k
+ theta.k<-theta.kp1
+ cat(it, theta.kp1, "\n")
+ }
1 0.938414
2 1.631730
3 1.933344
4 2.034420
5 2.065981
6 2.075641
7 2.078580
8 2.079472
9 2.079743
```

```
10 2.079826
11 2.079851
12 2.079858
13 2.079860
14 2.079861
```

Παρατηρείται λοιπόν ότι η σύγκλιση στην εκτίμηση του θ επέρχεται μετά από δεκατέσσερις επαναλήψεις, δίνοντας την τιμή `2.079861` με ακρίβεια στο έκτο δεκαδικό στοιχείο.

23.3 Τρίτο Παράδειγμα: Εκτίμηση Κανονικού Μοντέλου Πεπερασμένης Μίξης

Ένα κανονικό μοντέλο μίξης μπορεί να ορισθεί από δύο κανονικές κατανομές, $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$. Η πιθανότητα μία τυχαία μεταβλητή (αυτή που μπορεί να παρατηρηθεί) να ακολουθεί την πρώτη κατανομή είναι w. Η παράμετρος σε αυτό το μοντέλο είναι το διάνυσμα $\theta = (w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. Η συνάρτηση πυκνότητας πιθανότητας της μίξης δίνεται από:

$$p(y;\theta) = wp_1(y;\mu_1,\sigma_1^2) + (1-w)p_2(y;\mu_2,\sigma_2^2),$$

όπου $p_j(y;\mu_j,\sigma_j^2)$ είναι η συνάρτηση πυκνότητας πιθανότητας της κανονικής με παραμέτρους μ_j και σ_j^2.

Στον τυπικό μετασχηματισμό $C = (X, U)$, το X συμβολίζει τα δεδομένα που έχουν παρατηρηθεί, και το U δίνει την κατηγορία των δεδομένων που δεν έχουν παρατηρηθεί. Έστω $U = 1$ αν η παρατήρηση είναι από την πρώτη κατανομή, και $U = 0$ αν η παρατήρηση είναι από την δεύτερη κατανομή. Η μη δεσμευμένη αναμενόμενη τιμή $E(U)$ δίνει την πιθανότητα μία παρατήρηση να προέρχεται από την πρώτη κατανομή, η οποία είναι ίση με w.

Έστω n παρατηρήσεις του X, $x_1, \ldots, x_n$. Δεδομένης μιας αρχικής τιμής του θ, μπορεί να εκτιμηθεί η δεσμευμένη αναμενόμενη τιμή $E(U/x)$ για οποιαδήποτε πραγμάτωση του X:

$$E(U/x, \theta^{(k)}) = \frac{w^{(k)} p_1(x;\mu_1^{(k)}, \sigma_1^{2(k)})}{p(x; w^{(k)}, \mu_1^{(k)}, \sigma_1^{2(k)}, \mu_2^{(k)}, \sigma_2^{2(k)})}.$$

Το βήμα Μ του Ε-Μ αλγόριθμου είναι οι γνωστές Ε.Μ.Π. των παραμέτρων:

$$w^{(k+1)} = \frac{1}{n} \sum E(U|x_i, \theta^{(k)})$$

$$\mu_1^{(k+1)} = \frac{1}{nw^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)}) x_i$$

$$\sigma_1^{2(k+1)} = \frac{1}{nw^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_1^{(k+1)})^2$$

$$\mu_2^{(k+1)} = \frac{1}{n(1-w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)}) x_i$$

$$\sigma_2^{2(k+1)} = \frac{1}{n(1-w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_2^{(k+1)})^2$$

Για να δούμε πως δουλεύει ο αλγόριθμος για την εκτίμηση της w, παράγουμε μερικά τεχνητά δεδομένα στην R:

```
> # Normal mixture. Generate data from normal mixture with w=0.7,
> # mu_1=0, sigma^2_1=1, mu_2=1, sigma^2_2=2.
> # Note that R uses sigma, rather than sigma^2 in rnorm.
> # Set the seed, so computations are reproducible.
>
> set.seed(4)
> n<-300
> w<-0.7
> mu1<-0
> sigma21<-1
> mu2<-5
> sigma22<-2
> x<-ifelse(runif(n)<w, rnorm(n,mu1,sqrt(sigma21)),rnorm(n,mu2,sqrt(sigma22)))

> # Initialize
> theta.k<-.1
> theta.kp1<-.5
>
> it  <- 0  ####iterative count
> del <- 1  ####iterative adjustment
> while(abs(del) > 0.000001 && (it <- it+1) < 20) ###Loop for 20 iterations
+                                                 ###and prespecified presicion
+ {
+ tmp<-theta.k*dnorm(x,mu1,sqrt(sigma21))
+ ehat.k<-tmp/(tmp+(1-theta.k)*dnorm(x,mu2,sqrt(sigma22)))
```

```
+ theta.kp1<-mean(ehat.k)
+ del<- theta.kp1-theta.k
+ theta.k<-theta.kp1
+ cat(it, theta.kp1, "\n")
+ }
1 0.6130451
2 0.6686083
3 0.6715901
4 0.671751
5 0.6717596
6 0.6717601
```

Όπως φαίνεται από τα αποτελέσματα, ο αλγόριθμος συγκλίνει σε έξι επαναλήψεις, με ακρίβεια στο έκτο δεκαδικό στοιχείο, στην εκτίμηση `0.6717601`.