Ο αλγόριθμος ΕΜ

Ο αλγόριθμος ΕΜ αποτελεί μία αριθμητική μέθοδο που χρησιμοποιείται για την εύρεση εκτιμητριών ΜΠ για συγκεκριμένα προβλήματα. Επειδή έχει κατασκευαστεί με στατιστικές τεχνικές προσφέρει αρκετή πληροφορία στην επίλυση στατιστικών προβλημάτων. Ο αλγόριθμος ΕΜ μελετήθηκε διεξοδικά από τους *Dempster* (1977). Ο αλγόριθμος προϋπήρχε σε διάφορες μορφές πριν από αυτή τη χρονολογία (πρωτοεμφανίστηκε στις αρχές του 1900).

Ο αλγόριθμος ΕΜ: η βασική ιδέα

Η βασική ιδέα είναι να αυξήσουμε τα παρατηρούμενα δεδομένα με κάποια μη παρατηρούμενα δεδομένα (δεδομένα που λείπουν *missing*). Όταν αναφερόμαστε σε δεδομένα *missing* δεν σημαίνει απαραίτητα ότι είναι *missing* με την κλασική έννοια, αλλά μπορούμε να τα θεωρήσουμε σαν *missing* για την ευκολία της εκτίμησης. Αυτή η διαδικασία (να αυξήσουμε τα παρατηρούμενα δεδομένα με κάποια *missing* δεδομένα) ονομάζεται *data augmentation*.

Παραδείγματα από ¨*missing data*¨ I

- *Missing Data* (κάποιες μεταβλητές για κάποιες παρατηρήσεις δεν έχουν παρατηρηθεί. Δηλαδή ο πίνακας δεδομένων δεν είναι πλήρης.)
- Λογοκριμένες παρατηρήσεις (π.χ. στην ανάλυση επιβίωσης για κάποιους ασθενείς γνωρίζουμε ότι επιβίωσαν μέχρι κάποια χρονική στιγμή αλλά δεν γνωρίζουμε κάτι για μετά (είτε γιατί τέλειωσε η έρευνα, είτε γιατί τα άτομα αποσύρθηκαν από αυτήν)
- Περικομμένες παρατηρήσεις (κάποιες παρατηρήσεις δεν μπορούν να παρατηρηθούν λόγο περιορισμών στο σχέδιο δειγματοληψίας)
- Ομαδοποιημένα δεδομένα (γνωρίζουμε το διάστημα μέσα στο οποιό πέφτει κάποια παρατήρηση αλλά δεν γνωρίζουμε την ακριβή τιμή της παρατήρησης. π.χ. έρευνες με χρήση ερωτηματογίων).

Παραδείγματα από ¨*missing data*¨ II

- Μείξεις (Πολλά μοντέλα και πολλές κατανομές προκύπτουν από απλούστερα μοντέλα (κατανομές) μέσα από τη διαδικασία της μείξης.)
- Συνελίξεις κατανομών (*Convolutions*) (η τ.μ. που παρατηρούμε είναι το άθροισμα δύο επιμέρους τ.μ., όμως εμείς παρατηρούμε μόνο το άθροισμα και όχι τις επιμέρους τιμές)
- Τυχαία αθροίσματα (παρατηρούμε μια τυχαία μεταβλητή που προκύπτει ως άθροισμα ισόνομων τυχαίων μεταβλητών (τις οποίες δεν παρατηρούμε) των οποίων όμως το πλήθος δεν είναι σταθερό αλλά τυχαία μεταβλητή που ακολουθεί κάποια διακριτή κατανομή)
- Λανθάνοντα (*Hidden*) Μοντέλα Μάρκοβ (μοντέλα για εξαρτημένα δεδομένα. Παρατηρούμε μια χρονολογική σειρά, η τιμή όμως κάθε χρονική στιγμή εξαρτάται από μια μη παρατηρούμενη κατάσταση (*state*). Παρατηρούμε μόνο την τιμή και όχι τις καταστάσεις σε κάθε χρονική στιγμή).

EM: Η ιδέα

Έστω ότι θέλουμε να μεγιστοποιήσουμε την πιθανοφάνεια $L(\theta \mid Y)$, όπου Y είναι τα παρατηρούμενα δεδομένα. Αυξάνουμε τα παρατηρούμενα δεδομένα Y με επιπρόσθετα (μη παρατηρούμενα) δεδομένα Z έτσι ώστε η πιθανοφάνεια $L(\theta \mid Y, Z)$ να είναι ευκολότερο να μεγιστοποιηθεί. Ο αλγόριθμος προχωράει εκτιμώντας τα Z από την τρέχουσα εκτίμηση της $L(\theta \mid Y)$ (*E-step*) και μεγιστοποίηση της $L(\theta \mid Y, Z)$ ως προς θ χρησιμοποιώντας για Z τις τιμές που προέκυψαν από το *E-step*. Μπορεί να αποδειχθεί ότι η πιθανοφάνεια αυξάνεται σε κάθε επανάληψη.

EM: πως δουλεύει

Δοθέντος μίας εκτίμησης για το $\theta^{(r)}$ ορίζουμε τη συνάρτηση

$$\begin{aligned} Q(\theta, \theta^{(r)}) &= \int_Z \log L(\theta \mid Y, Z) P(Z \mid \theta^{(r)}, Y) dZ \\ &= E(\log L(\theta \mid Y, Z)) \end{aligned}$$

και $P(Z \mid \theta^{(r)}, Y)$ είναι η "*posterior*" κατανομή του Z δοθέντος της τρέχουσας εκτίμησης και των δεδομένων. Ο αλγόριθμος EM ορίζεται ως:
Δοθέντος της παρούσας εκτίμησης $\theta^{(r)}$

1. *E-step* Υπολόγισε $Q(\theta, \theta^{(r)})$
2. *M-step* Μεγιστοποίησε $Q(\theta, \theta^{(r)})$ ως προς θ.

EM ; Πλεονεκτήματα και μειονεκτήματα

Πλεονεκτήματα

- Εκτιμήσεις σε αποδεκτό εύρος όταν οι αρχικές τιμές είναι σε αποδεκτό εύρος.
- Εύκολος προγραμματιστικά
- Ενδιαφέρουσα Στατιστική ερμηνεία
- Υπο-προϊόντα του αλγόριθμου είναι χρήσιμα για περαιτέρω στατιστική συμπερασματολογία.

Μειονεκτήματα

- Αργή σύγκλιση (ποιο αργή από *Newton-Raphson* για παράδειγμα)
- Εντοπισμός τοπικών αντί για ολικά μέγιστα (απαιτείται η χρήση πολλών αρχικών τιμών)
- Η λύση εξαρτάται από τις αρχικές τιμές.

Χρήσιμα στοιχεία

- Συνήθως ο αλγόριθμος χρειάζεται μόνο ένα μικρό αριθμό επαναλήψεων για να προσεγγίσει τη λύση αλλά όταν είναι κοντά στη λύση μπορεί να χρειαστεί πολύς χρόνος για να εντοπίσει τη λύση. Μία χρήσιμη ιδέα είναι να χρησιμοποιηθεί αρχικά ο αλγόριθμος ΕΜ για την προσέγγιση της λύσης και μετά κάποιος άλλος αλγόριθμος όπως ο *Newton-Raphson* για τον εντοπισμό της
- Η πιθανοφάνεια μπορεί να συνεχίσει να αυξάνεται ενώ οι παράμετροι μπορεί να φαίνεται ότι έχουν σταθεροποιηθεί. Το κριτήριο τερματισμού είναι σημαντικό.
- Ο ρυθμός σύγκλισης εξαρτάται στην πληροφορία που λείπει. Αν η πληροφορία που λείπει είναι μεγάλη τότε ο αλγόριθμος μπορεί να είναι πολύ αργός.

Συνθήκες τερματισμού

- Βάση της μεταβολής στην πιθανοφάνεια
  Σταμάτα τις επαναλήψεις όταν:

$$\left| \frac{L^{(r+1)} - L^{(r)}}{L^{(r+1)}} \right| \leq tol$$

- Βάση των μεταβολών στις παραμέτρους
  Σταμάτα τις επαναλήψεις όταν:

$$max_j \left(|\theta_j^{(r+1)} - \theta_j^{(r)}| \right) \leq tol$$

Και οι 2 συνθήκες υποδεικνύουν έλλειψη προόδου παρά σύγκλιση!

Ο αλγόριθμος ΕΜ στην εκθετική οικογένεια

Εάν η πιθανοφάνεια για τα πλήρη δεδομένα ανήκει στη εκθετική οικογένεια τότε για τη μεγιστοποίηση στο *M-step* απαιτούνται μόνο επαρκείς στατιστικές ποσότητες. Σε αυτή την περίπτωση το *E-step* εμπλέκει μόνο:

- Απλές αναμενόμενες τιμές (αναμενόμενες τιμές απλών συναρτήσεων)
- Στο *E-step* δεν χρειαζόμαστε την αναμενόμενη τιμή της κάθε παρατήρησης αλλά την αναμενόμενη τιμή μιάς (επαρκής) στατιστικής ποσότητας το οποίο είναι ποιο απλό και γρήγορο.

Τυπικά σφάλματα

Μιας και ο ΕΜ αποφεύγει τον υπολογισμό των δεύτερων παραγώγων τα τυπικά σφάλματα δεν είναι διαθέσιμα μετά τον τερματισμό του αλγορίθμου. Μία λύση για αυτό το ζήτημα μπορεί να είναι η "*missing information principle*":
Observed Information = Complete Information - Missing Information
Αυτό γράφεται ως:

$$-\frac{d^2\ell(\theta \mid Y)}{d\theta_i d\theta_j} = \left[-\frac{d^2Q(\theta,\phi)}{d\theta_i d\theta_j}\right]_{\phi=\theta} - \left[-\frac{d^2H(\theta,\phi)}{d\theta_i d\theta_j}\right]_{\phi=\theta}$$

Χωρίς τα *missing data* ο πρώτος όρος στη δεξιά πλευρά θα ήταν ο πίνακας πληροφορίας. Ένα χρήσιμο αποτέλεσμα είναι το ακόλουθο:

$$-\left[\frac{d^2H(\theta,\phi)}{d\theta_i d\theta_j}\right] = Var\left[\frac{d\ell(\theta \mid Y,Z)}{d\theta}\right]$$

Υπάρχουν και άλλες μεθοδολογίες για την εξαγωγή τυπικών σφαλμάτων από την ιστορία του αλγόριθμου ΕΜ (παρόλο που δεν θα τις συζητήσουμε εδώ).

Παράδειγμα 1:

Το παράδειγμα αφορά τη γενετική σύνδεση 197 ζώων. Τα ζώα κατανέμονται σε 4 κατηγορίες:

$$Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

με πιθανότητες κελιών:

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right)$$

Θέλουμε να βρούμε εκτιμήσεις μέγιστης πιθανοφάνειας για το θ. Αριθμητικές μέθοδοι απαιτούνται για τη μεγιστοποίηση της πιθανοφάνειας.

Παράδειγμα 1: Αύξηση των δεδομένων

Αυξάνουμε τα παρατηρούμενα δεδομένα Y διαιρώντας το πρώτο κελί σε δύο κελιά. Έστω $X = (x_0, x_1, x_2, x_3, x_4)$ τα αυξημένα δεδομένα, με $x_i = y_i$, $i = 2, 3, 4$ και $x_0 + x_1 = y_1$, όπου οι πιθανότητες για το x_0 και το x_1 είναι $1/2$ και $\theta/4$, αντίστοιχα.
Η παρατηρούμενη πιθανοφάνεια είναι:

$$L(\theta \mid Y) \propto (2 + \theta)^{y_1}(1 - \theta)^{y_2+y_3}\theta^{y_4}$$

ενώ η πλήρης πιθανοφάνεια είναι:

$$L(\theta \mid X) \propto (1 - \theta)^{x_2+x_3}\theta^{x_4+x_1}$$

η οποία είναι πολύ ποιο εύκολη. Επιπλέον,
$x_1 \mid \theta, Y \sim \mathit{Binomial}(125, \frac{\theta}{\theta+2})$ (χρησιμοποιούμε το γεγονός ότι: $\frac{\theta}{\theta+2} = \frac{\theta/4}{\theta/4+1/2}$).

Παράδειγμα 1: Ο αλγόριθμος EM

E-step: Υπολόγισε

$$E(x_1 \mid \theta, Y) = \frac{125\theta}{\theta + 2} = t$$

M-step: Πάρε νέα τιμή για το θ ως:

$$\theta^{(new)} = \frac{t + x_4}{t + x_2 + x_3 + x_4}$$

Στο παράδειγμα μας ξεκινώντας από $\theta^{(0)} = 0.40$ ο αλγόριθμος συγκλίνει μετά από 8 επαναλήψεις με κριτήριο τερματισμού αν η διαφορά ανάμεσα σε δύο διαδοχικές εκτιμήσεις που παίρνουμε είναι μικρότερη από 10^{-6}. Οι διαδοχικές τιμές για τις εκτιμήσεις είναι:

$$(0.4, 0.5906643, 0.6218892, 0.6261642, 0.6267342,$$

$$0.6268099, 0.626820, 0.6268213, 0.6268215).$$

Παράδειγμα 1 : Ο αλγόριθμος EM : Περισσότερες λεπτομέρειες

Από τη θεωρία, το *E-step* είναι απλά ο υπολογισμός του

$$\begin{aligned} Q(\theta, \theta^{(r)}) &= \int_Z \log L(\theta \mid Y, Z) P(Z \mid \theta^{(r)}, Y) dZ = E(\log L(\theta \mid Y, Z)) \\ &= constant + E\left[(x_2 + x_3)\log(1-\theta) + (x_1 + x_4)\log\theta\right] \\ &= constant + (x_2 + x_3)\log(1-\theta) + (E\left[x_1\right] + x_4)\log\theta \end{aligned}$$

η αναμενόμενη τιμή υπολογίζεται ως προς την δεσμευμένη κατανομή του $x_1 \mid Y, \theta^{(r)}$.

Στο *M-step* αντικαθιστούμε την αναμενόμενη τιμή από το *E-step* και μεγιστοποιούμε σα συνάρτηση του θ. Αυτό αντιστοιχεί σε μεγιστοποίηση της λογαριθμοπιθανοφάνειας των πλήρη δεδομένων, όπου τα μη παρατηρούμενα δεδομένα αντικαθιστούνται από τις αναμενόμενες τιμές τους (συγκεκριμένα, οι *συναρτήσεις* που εμπλέκονται στην λογαριθμοπιθανοφάνεια των πλήρη δεδομένων εκτιμώνται από τις αναμενόμενες τιμές τους.)

Παράδειγμα 1 : ο αλγόριθμος *NR*

Για αυτό το παράδειγμα ο αλγόριθμος *NR* εκτελεί τις επαναλήψεις χρησιμοποιώντας:

$$\theta^{(new)} = \theta - \frac{\frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta}}{-\frac{y_1}{(2+\theta)^2} - \frac{y_2+y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2}}$$

Ξεκινώντας από τις ίδιες αρχικές τιμές ο αλγόριθμος χρειάζεται 4 επαναλήψεις προκειμένου να καταλήξει στην ίδια τιμή. Οι διαδοχικές τιμές είναι:

$$(0.6170669, 0.6269629, 0.6268215, 0.6268215)$$

Παράδειγμα 1 : ο αλγόριθμος *NR*: περισσότερες λεπτομέρειες

- Παρατηρούμενος *Fisher-Information matrix*:

$$J(\theta) = \frac{y_1}{(2+\theta)^2} + \frac{y_2 + y_3}{(1-\theta)^2} + \frac{y_4}{\theta^2}$$

- Αναμενόμενος *Fisher Information matrix*:

$$I(\theta) = \frac{n}{4(2+\theta)^2} + \frac{2n}{4(1-\theta)^2} + \frac{n}{4\theta^2}$$

Η *updated* τιμή της παραμέτρου δίνεται από τη σχέση:

$$\theta^{(new)} = \theta - \frac{\frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta}}{-\frac{n}{4(2+\theta)} - \frac{2n}{4(1-\theta)^2} - \frac{n}{4\theta^2}}$$