

EM ALGORITEM

Pogosto srečamo v statistiki problem manjkajočih opazovanih vrednosti. Obstaja mnogo metod, kako korektno oceniti parametre. Ogleдали si bomo poseben primer EM (expectation maximization) algoritma, ki je eden od pristopov.

- a. Prepostavite, da so vaše opazovane vrednosti neodvisni p -razsežni normalni vektorji $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ s parametri μ in Σ . Ocenite parametra po metodi največjega verjetja, če ni manjkajočih podatkov.
- b. Predpostavite, da nekatere komponente “opazovanih” vektorjev manjkajo. Prepostavite, da so podatki manjkajo “naključno” in neodvisno od $\mathbf{X}_1, \dots, \mathbf{X}_n$, vendar tako, da nikoli ne manjkajo vse komponente. Označimo z $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ opazovane vrednosti (z manjkajočimi podatki).

EM algoritem ima dva koraka:

- (i) E-KORAK: Naj bo $\ell_c(\mu, \Sigma | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ logaritemska funkcija verjetja, če imamo vse podatke. Indeks c pomeni “complete”. Te funkcije ne moremo izračunati, če kakšen podatek manjka. Kaj storiti? Označimo z $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ dejansko opazovane “škrbaste” vektorje. Izberimo začetni približek za parametra μ in Σ , recimo μ_0 in Σ_0 . Izračunajmo pogojno matematično upanje

$$Q(\mu, \Sigma, \mu_0, \Sigma_0) = E\left(\ell_c(\mu, \Sigma | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) | \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\right).$$

Pri tem privzemamo, da so $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ porazdeljeni večrazsežno normalno s parametroma μ_0 in Σ_0 .

- (ii) M-KORAK: Naslednja približka μ_1 in Σ_1 za neznan parametra dobimo tako, da maksimiziramo funkcijo $Q(\mu, \Sigma, \mu_0, \Sigma_0)$ po μ in Σ .

Koraka E in M potem ponavljamo. Ponovimo E-korak z novimi približki za parameter in “pridelamo” nove približke z M-korakom. V mnogo primerih (glej Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B*,

39, 1-38) zaporedni približki konvergirajo proti neki limiti, ki je potem naša ocena za parametre.

Na kratko komentirajte, kaj mislite o tem postopku? Se vam zdi smiseln? Zakaj?

- c. Opišite, s čim se nadomestijo manjkajoče vrednosti v primeru večrazsežne normalne porazdelitve. Utemeljite vaše izjave. Lahko se omejite na primer $p = 2$. Kako smiseln se vam zdi zdaj EM algoritem? Na kratko komentirajte.
- d. Naj bo $p = 2$. Generirajte vzorec velikosti $n = 400$. Za vsak $k = 1, 2, \dots, n$ naj manjka ena od komponent z verjetnostjo $0, 1$ in sicer manjkajoči podatek izberite naključno z verjetnostjo $1/2$. Sprogramirajte EM algoritem in ugotovite, ali zaporedni približki res konvergirajo. Primerjajte limitne ocene s tistimi, ki bi jih dobili z metodo največjega verjetja samo na podlagi podatkov, kjer ne manjka nobena komponenta. Komentar?

Literatura: Geoffrey J. McLachlan, Thriyambakam Krishnan, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, 1997.