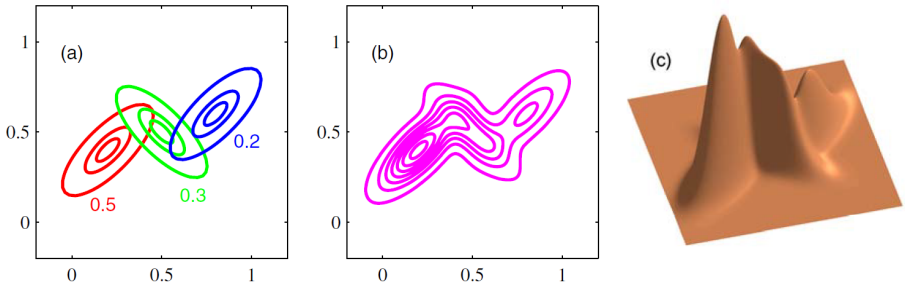
Mladen Nikolić Anđelka Zečević

MAŠINSKO UČENJE

Beograd
2019.



Slika 14.4: Ilustracija mešavine normalnih raspodela. Na slici (a) prikazane su konture tri normalne raspodele sa pridruženim verovatnoćama da tačka pripada klasteru (brojevima 0.5, 0.3 i 0.2). Na slici (b) prikazane su konture mešavine, a na slici (c) njen grafik.

14.2 Mešavina normalnih raspodela i EM algoritam

Mešavina normalnih raspodela (eng. *mixture of Gaussians*) predstavlja nešto sofisticiraniji model klasterovanja od modela k sredina, ali je zanimljivo primetiti da postoji i veza, koja će biti objašnjena na kraju. Osnovna pretpostavka je da se podaci mogu podeliti u određeni broj relativno kompaktnih globularnih klastera čiji se oblik može dobro opisati normalnim raspodelama sa različitim prosecima i matricama kovarijacije. Proseci, jasno, definišu pozicije klastera u prostoru, dok matrice kovarijacije opisuju njihov oblik i orijentaciju u prostoru. Površi jednake gustine u okviru jednog klastera u tom slučaju predstavljaju elipsoide. U ovom modelu, raspodela je malo složenija nego u dosadašnjim, ali ima prirodnu dekompoziciju na jednostavnije raspodele. Zamislimo kako se može dobiti nasumično generisana tačka iz ovakve raspodele. Prvo se može nasumice izabrati klaster, a potom se iz tog klastera može izabrati tačka u skladu sa normalnom raspodelom koja mu odgovara. Ukoliko je C broj klastera i za brojeve $p_1, \ldots, p_C$ važi $p_i \geq 0$ i $\sum_{i=1}^{C} p_i = 1$, onda $(p_1, \ldots, p_C)$ predstavlja multinomijalnu raspodelu nad klasterima. Gustina raspodele nad instancama se može zapisati kao:

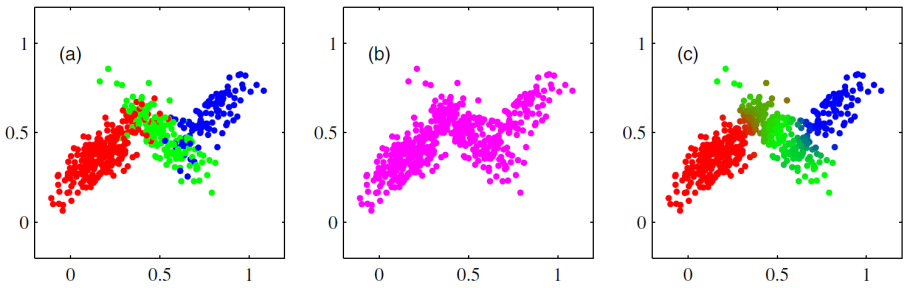
$$p(x) = \sum_{i=1}^{C} p_i \mathcal{N}(x; \mu_i, \Sigma_i)$$

gde je $\mathcal{N}$ gustina normalne raspodele sa više promenljivih

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Ovo je ilustrovano slikom 14.4.

Primetimo da ovaj model prestavlja generativni model i mogao bi se koristiti u različite svrhe, pa i u svrhe klasifikacije. Pošto svakoj klasi može odgovarati



Slika 14.5: Ilustracija podataka iz mešavine normalnih raspodela. Na slici (a) prikazani su podaci kao uzorci iz raspodele $p(x, z) = p(x|z)p(z)$, zbog čega se na slici vidi i njihova boja, koja se dobija na osnovu multinomijalne raspodele i pozicija, koja se dobija na osnovu normalne raspodele. Na slici (b), prikazana je informacija koja se može dobiti na osnovu marginalne raspodele $p(x)$ i otud nije poznata informacija o promenljivim z, odnosno o bojama. Na slici (c) tačke su obojene mešavinom boja klastera, pri čemu je svaka od boja zastupljena proporcionalno verovatnoći da tačka pripada tom klasteru.

jedan klaster, uslovna raspodela $p(x|z = c)$ je definisana kao dogovarajuća normalna raspodela, a ona se može koristiti za klasifikacju, jer Bajesova teorema omogućava da se na osnovu nje izračuna $p(y|x)$. Ipak, time se nećemo baviti. U kontekstu klasterovanja, neka promenljiva z uzima vrednosti od 1 do C u skladu sa multinomijalnom raspodelom $(p_1, \ldots, p_C)$. Verovatnoća da instanca pripada klasteru k je

$$p_w(z = k|x) = \frac{p_w(x|z = k)p_w(z = k)}{p_w(x)} = \frac{p_w(x|z = k)p_w(z = k)}{\sum_{k=1}^{C} p_w(x|z = k)p_w(z = k)}$$

gde je w vektor parametara koji objedinjuje sve parametre iz svih vektora μ_k i matrica Σ_k, kao i $(p_1, \ldots, p_C)$. Ilustracija podataka kao uzoraka iz mešavine raspodela, data je na slici 14.5

Primetimo da, iako se koriste u modelu, vrednosti promenljive z za različite instance nisu poznate u vreme obučavanja. Primetimo takođe, da bi problem bio vrlo jednostavan ukoliko bi bile. Kada su poznate vrednosti z_i za svaku od instanci, parametri μ_k i Σ_k normalne raspodele klastera k se ocenjuju standardnim empirijskim ocenama – prosekom instanci za koje važi $z_i = k$ matricom kovarijacija između svih atributa izračunatom na osnovu tih instanci. U ovakvim situacijama, kada se može pretpostaviti da postoje takozvane *latentne*, odnosno *skrivene* promenljive z koje nisu opažene i kada je ocena parametara raspodele $p_w(x)$ teška, a ocena parametara raspodele $p_w(x, z)$ laka (kada su dati empirijski podaci i za x i za z), pribegava se algoritmu *maksimizacije očekivanja* (eng. *expectation maximization*) iliti, skraćeno – *EM algoritmu*.

EM algoritam ćemo prikazati prvo u opštem obliku, pa ćemo ga precizirati

za model mešavine normalnih raspodela. Za logaritam funkcije verodostojnosti parametara važi:

$$\ell(w) = \log \mathcal{L}(w) = \sum_{i=1}^{N} \log p_w(x_i) = \sum_{i=1}^{N} \log \int_z p_w(x_i, z) dz$$

ili u diskretnom slučaju

$$\ell(w) = \log \mathcal{L}(w) = \sum_{i=1}^{N} \log p_w(x_i) = \sum_{i=1}^{N} \log \sum_{k=1}^{C} p_w(x_i, k)$$

Ovaj problem nije lak za rešavanje, pošto logaritam ne može da prođe kroz sumu. Rešenje se sastoji u posmatranju funkcije verodostojnosti u odnosu na zajedničku raspodelu promenljivih x i z. Imajući u vidu da svakoj instanci pored vrednosti opaženih promenljivih x odgovaraju i vrednosti latentnih promenljivih z, ovo je potpuno legitimno, ali kako vrednosti z nisu poznate, takva funkcija verodostojnosti se ne može izračunati, već predstavlja slučajnu promenljivu. Iako se ona ne može izračunati, može se izračunati njeno očekivanje u odnosu na promenljive z, tako da optimizacija može biti vršena po njemu. Konkretno, logaritam funkcije veodostojnosti koja uzima u obzir i promenljive z je

$$\ell(w) = \sum_{i=1}^{N} \log p_w(x_i, z_i)$$

Neka je Q pomoćna funkcija

$$Q(w, w^{t-1}) = \mathbb{E}_z[\ell(w)|w^{t-1}] = \int \ell(w) p_{w^{t-1}}(z) dz$$

Nova vrednost parametara w se dobija maksimizacijom funkcije Q. Konkretno

$$w^t = \arg\max_w Q(w, w^{t-1})$$

Ovako definisan algoritam se može podeliti u dva koraka. U prvom se ocenjuje funkcija Q. Ovaj korak se naziva E, pošto se u tom postupku zapravo vrši ocena očekivanja. U drugom koraku se vrši maksimizacija, pa se zato naziva korak M. Ovaj algoritam ima jedno zanimljivo svojstvo, a to je da se vrednost funkcije Q nikad ne smanjuje, što može biti korisno pri debagovanju njegove implementacije.

U slučaju problema klasterovanja pomoću mešavine normalnih raspodela,

funkcija Q ima sledeći oblik:

$$\begin{aligned}
Q(w, w^{t-1}) &= \mathbb{E}\left[\sum_{i=1}^{N} \log p_w(x_i, z_i)\right] \\
&= \sum_{i=1}^{N} \mathbb{E}\left[\log\left[\prod_{k=1}^{C} p_k p_w(x_i)^{I(z_i=k)}\right]\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{C} \mathbb{E}[I(z_i = k)] \log[p_k p_w(x_i)] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{C} p_{w^{t-1}}(z_i = k|x_i) \log[p_k p_w(x_i)] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{C} p_{w^{t-1}}(z_i = k|x_i) \log p_k + \sum_{i=1}^{N}\sum_{k=1}^{C} p_{w^{t-1}}(z_i = k|x_i) \log p_w(x_i)
\end{aligned}$$

Korak E se sastoji od izračunavanja veličine $p_{w^{t-1}}(z_i = k|x_i)$ na sledeći način:

$$p_{w^{t-1}}(z_i = k|x_i) = \frac{p_k \mathcal{N}(x; \mu_k^{t-1}, \Sigma_k^{t-1})}{\sum_{j=1}^{C} p_j \mathcal{N}(x; \mu_j^{t-1}, \Sigma_j^{t-1})}$$

Kako su sve vrednosti koje izraz uključuje definisane u prethodnom koraku, ovaj izraz je lako izračunati.

Korak M se sastoji od maksimizacije funkcije Q po w, odnosno po veličinama p_k, μ_k i Σ_k. Može se izvesti sledeće rešenje:

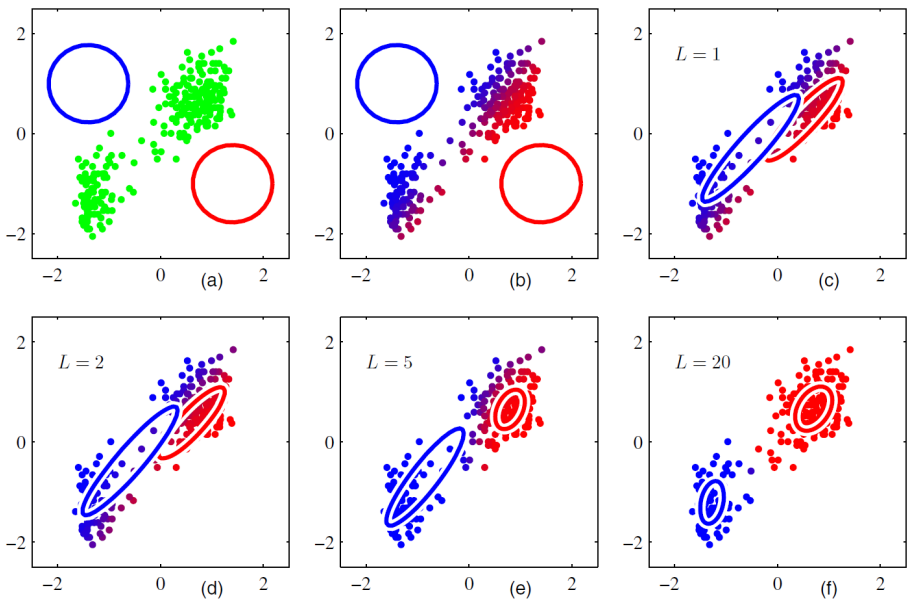
$$p_k = \frac{1}{N}\sum_{i=1}^{N} p_{w^{t-1}}(z_i = k|x_i)$$

$$\mu_k = \frac{\sum_{i=1}^{N} p_{w^{t-1}}(z_i = k|x_i) x_i}{\sum_{i=1}^{N} p_{w^{t-1}}(z_i = k|x_i)}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} p_{w^{t-1}}(z_i = k|x_i)(x_i - \mu_k)^T (x_i - \mu_k)}{\sum_{i=1}^{N} p_{w^{t-1}}(z_i = k|x_i)}$$

Primetimo da su izvedene formule intuitivne. Naime, p_k je praktično udeo instanci koje su pridružene klasteru i, s tom ogradom da sad svaka instanca pripada svakom klasteru, ali sa težinom $p_{w^{t-1}}(z_i = k|x_i)$. Prosek ovih težina za klaster k je ocena p_k. Slično, prosek klastera k je prosek svih instanci, ali otežanih u skladu sa udelom sa kojim pripadaju tom klasteru. Interpretacija formule za kovarijansu je potpuno analogna. Izvršavanje ovog algoritma u više iteracija ilustrovano je na slici 14.6.

Primetimo da se algoritam k sredina može posmatrati kao jednostavnija verzija klasterovanja pomoću mešavine normalnih raspodela koja se karakteriše sledećim pretpostavkama:



Slika 14.6: Izvršavanje EM algoritma za mešavinu normalnih raspodela u 20 koraka.

- $\Sigma_k = \sigma^2 I$ za svako k,
- $p_k = 1/C$ za svako k i
- $p(z_i = k|x_i)$ je 1 ako je među svim prosecima μ_j instanci x_i najbliži baš μ_k, a 0 u suprotnom

Očito, klasterovanje zasnovano na mešavini normalnih raspodela je značajno izražajnije – može prirodno modelovati ne samo loptaste klastere, već i klastere u obliku elipsoida različitih orijentacija u prostoru. Pritom njihov oblik se može ustanoviti na osnovu matrice kovarijanse, što znači da ovaj model omogućava i interpretaciju. Takođe, dodele instanci klasterima nisu tvrde, već mogu izražavati i pouzdanost da instanca pripada nekom klasteru.