Rješenje zadatka 5.3 predmeta Strojno učenje

Siniša Biđin

6. veljače 2013.

(a) Algoritam je implementiran programskim jezikom Octave, a nalazi se u datotekama `main.m, kmeans.m, kriterijska.m` i `a3.m`.

Korištene su značajke 2, 6, 9, 11 i 15 (gdje je prva značajka 1, a ne 0).

(b) (i) Za svaki K zapisujemo broj iteracija potrebnih do konvergencije i vrijednost kriterijske funkcije J za ono od 10 grupiranja u kojem je vrijednost J bila najmanja.

K	final iter	J
2	13	205648.875
3	13	205648.875
4	7	146690.680
5	11	146690.680
6	10	146690.680
7	9	146690.680
8	7	146690.680

(ii) Grafički prikažemo ovisnost kriterijske funkcije o broju grupa te tražimo "koljeno" krivulje. Vrijednost funkcije pada porastom K, no pad prestaje biti značajan nakon što K postane dovoljno velik i algoritam kreće razdijeljivati prirodne grupe.

Na temelju rezultata iz (i) odlučili bismo se za idealan $K = 4$, pošto nakon njega vrijednost kriterijske funkcije značajno ne pada. To je ujedno i stvarni broj klasa u skupu podataka.

(iii) Uz najbolje slučajno odabrane početne centroide iz `k4centroids.mat`, započinjemo grupiranje u $K = 4$ klase. Centroidi konvergiraju nakon 7 iteracija.

iter	J
1	300320.131
2	197605.057
3	182869.890
4	149302.633
5	146928.357
6	146690.680
7	146690.680

Zbirni rezultati grupiranja su sljedeći:

G_1: bus 30, saab 95, van 3, opel 105
G_2: *nijedan primjer*
G_3: bus 20, van 2
G_4: bus 116, saab 73, van 147, opel 66

Grupa G_2 ne sadrži nijedan primjer jer se centroid te grupe u ovom pokretanju algoritma inicijalizirao na vrijednost dovoljno daleku od svih primjera. To je moguće pošto su centroidi inicijalizirani na način da se srednjoj vrijednosti svih primjera nadodaju nasumični vektori. Vjerojatnost takvog rezultata smanjuje se boljim odabirom raspona mogućih vrijednosti nasumičnih vektora korištenih za inicijalizaciju centroida.

(c) Implementacija se nalazi u datotekama `em.m`, `normaliziraj.m`, `kaos.m`, `logizglednost.m`, `multigauss.m` i `trecidii.m`.

Inicijalizacijske vrijednosti parametara:

- Središta grupa postavljaju se na nasumične različite primjere u skupu.
- Svi koeficijenti mješavine postavljaju se na istu vrijednost.
- Kovarijacijska matrica procjenjuje se putem $\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}$.

(d) (i) Za svaki K ponavljamo 10 grupiranja i u tablicu dodajemo ona s najvećom vrijednošću log-izglednosti. Pronađeni parametri za $K = 4$ zapisani su u datoteku `k4em.txt`.

K	final iter	$\ln \mathcal{L}(\boldsymbol{\theta}\|\mathcal{D})$
2	6	710.735821
3	5	739.598856
4	3	726.324001
5	4	533.099234
6	4	551.657651
7	5	401.106718
8	4	317.867878

(ii) Središta učitavamo iz prethodnog zadatka. Koeficijente mješavine i kovarijacijske matrice inicijaliziramo na uobičajene, početne vrijednosti.

iteracija	$\ln \mathcal{L}(\boldsymbol{\theta}\|\mathcal{D})$
1	593.251514
2	696.332802
3	*konvergira*

Za svaku grupu ispisujemo, sortirane silazno prema vjerojatnosti, prvih par desetaka primjera koji najvjerojatnije pripadaju toj grupi i stvarne klase tih primjera. Pun popis nalazi se u datoteci `output.txt`. Imena klasa (bus, opel...) prevedena su u cijele brojeve 0, 1, 2 i 3.

```
Grupa 0:
   p         y
   -------   -------
   0.67983   1.00000
   0.65188   3.00000
   0.64599   1.00000
   0.64340   3.00000
   0.62613   3.00000
   0.62527   1.00000
   0.61278   1.00000
   0.61172   1.00000
   0.59826   1.00000
   0.58814   0.00000
   0.58701   1.00000
   0.57586   1.00000
   0.56612   3.00000
   0.55313   0.00000
   0.54371   3.00000
   0.54296   0.00000
   0.52768   1.00000
   0.51927   1.00000
   0.51458   0.00000
   0.51103   1.00000
   0.50587   3.00000
   0.50459   0.00000
   0.50231   0.00000
   ...
```

Grupa 1:

p	y
0.64973	3.00000
0.64966	3.00000
0.61913	3.00000
0.61750	3.00000
0.61550	1.00000
0.59590	3.00000
0.59213	3.00000
0.57675	1.00000
0.57669	3.00000
0.57075	1.00000
0.56926	1.00000
0.55071	3.00000
0.54885	3.00000
0.53300	3.00000
0.53290	1.00000
0.52841	1.00000
0.52490	1.00000
0.51687	3.00000
0.51540	3.00000
0.51285	3.00000
0.50913	3.00000
0.50875	3.00000
0.50704	3.00000
0.50463	3.00000
0.50358	3.00000
0.50181	1.00000
0.50107	3.00000
...	

Grupa 2:

p	y
1.00000	2.00000
1.00000	2.00000
1.00000	0.00000
1.00000	0.00000
1.00000	0.00000
1.00000	0.00000
1.00000	0.00000
1.00000	0.00000
1.00000	2.00000
1.00000	2.00000
1.00000	2.00000
0.92712	2.00000
0.88040	2.00000
0.87022	2.00000
0.84952	2.00000
0.81606	2.00000
0.80112	2.00000
0.76450	2.00000
0.76025	2.00000
0.75420	2.00000
0.69903	2.00000
0.63721	2.00000
0.58970	2.00000
0.56897	2.00000
0.54890	2.00000
0.54620	3.00000
0.47708	2.00000
0.42989	2.00000
0.41655	3.00000
0.40499	2.00000
...	

```
Grupa 3:
   p         y
   -------   -------
   0.42720   2.00000
   0.42271   2.00000
   0.41595   3.00000
   0.41399   1.00000
   0.41291   2.00000
   0.41145   2.00000
   0.40735   2.00000
   0.40727   2.00000
   0.40726   3.00000
   0.40720   1.00000
   0.40708   2.00000
   0.40533   3.00000
   0.40461   2.00000
   0.40412   2.00000
   0.40270   2.00000
   0.40130   1.00000
   ...
```

(e) Pokretanje algoritma koristeći središta dobivena u zadatku b-(iii) rezultira jednakim manjim brojem koraka potrebnih do konvergencije i sličnu konačnu log-izglednost. Manja razlika je takva vjerojatno zbog agresivnih (i neoptimalnih) provjera nastupanja konvergencije u ovoj implementaciji EM-algoritma: zaustavljam iteracije ranije nego što bilo najbolje.