Strojno učenje – domaća zadaća 5

UNIZG FER, ak. god. 2011/12.

Zadano: 18.1.2012. Rok predaje: 31.1.2012. do 17.00 sati.

Zadatak 1: Algoritam maksimizacije očekivanja

(a) Napišite funkciju (nepotpune) log-izglednosti za općenit model miješane gustoće. Što je problem s tom funkcijom?

(b) Definirajte E-korak i M-korak općenitog EM-algoritma. Zašto algoritam provodimo iterativno?

(c) Krenuvši od potpune log-izglednosti $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}, \mathcal{Z})$ za općenit model miješane gustoće, izvedite izraze za procjenu parametara π_k, $\boldsymbol{\mu}_k$ i $\boldsymbol{\Sigma}_k$ modela Gaussove mješavine.

(d) Na predavanjima smo spomenuli da je algoritam k-srednjih vrijednosti poseban slučaj EM-algoritma uz pretpostavku dijeljene izotropne kovarijacijske matrice. Uz tu pretpostavku, napišite izraz za potpunu log-izglednost $\ln \mathcal{L}(\theta|\mathcal{D}, \mathcal{Z})$. Usporedite dobiveni izraz s funkcijom pogreške J algoritma k-srednjih vrijednosti.

(e) Skicirajte krivulju log-izglednosti kao funkciju od broja iteracija EM-algoritma. Obrazložite izgled krivulje. Pretpostavite da smo algoritam pokretali još tri puta i na istome grafu ucrtajte odgovarajuće krivulje. Jesu li krivulje identične? Obrazložite odgovor.

(f) Skicirajte krivulju log-izglednosti kao funkciju broja grupa K. Obrazložite izgled krivulje. Možete li temeljem ove krivulje odrediti optimalan broj grupa? Kako?

Zadatak 2: Hijerarhijsko aglomerativno grupiranje

Zadan je skup primjera

$$\mathcal{D} = \left\{a = (0,1), b = (7,1), c = (2,1), d = (3,3), e = (2,4), f = (5,6), g = (9,10)\right\}.$$

(a) Skicirajte primjere. Nacrtajte dendrogram koji se dobiva hijerarhijskim aglomerativnim grupiranjem ovih primjera temeljem jednostruke povezanosti. Na dendrogramu naznačite udaljenost na kojoj je provedeno stapanje grupa.

(b) Nacrtajte dendogram koji se dobiva grupiranjem temeljem potpune povezanosti.

(c) Kako biste odredili optimalan broj grupa K temeljem dendrograma dobivenog u prethodnom zadatku? Obrazložite odgovor. Kojem pragu udaljenosti odgovara odabrani broj grupa? Presijecite dendogram na odabranom pragu udaljenosti i napišite particiju koja se time dobiva.

(d) Napišite matricu sličnosti za primjere iz $\mathcal{D}$. Pretpostavite da raspolažete samo tom matricom. Koje algoritme grupiranja (od onih koje smo upoznali na predavanjima) možemo primijeniti na takvu matricu?

Zadatak 3: Implementacija algoritama grupiranja

U ovom zadatku potrebno je implementirati algoritam k-srednjih vrijednosti i algoritam maksimizacije očekivanja te ih isprobati na vlastitom skupu podataka. *Napomena:* implementaciju (u bilo kojem programskom jeziku) priložite uz izvještaj.

(a) U ovom podzadatku potrebno je načiniti dvodimenzionalan skup od 500 primjera koji će se kasnije koristiti za isprobavanje algoritama za grupiranje. Odaberite nasumce 5 središta Gaussovih izvora iz skupa $[0, 100] \times [0, 100]$. Za svaki izvor nasumce odredite standardno odstupanje, tako da ono bude u rasponu $[10, 40]$.

Sada je potrebno iz svakog izvora uzorkovati po 100 primjera. U izvještaju je potrebno na jednoj slici prikazati sve uzorkovane primjere (tako da se primjeri iz različitih izvora razlikuju bilo po boji, bilo po obliku) i njihova središta. Zapišite uzorkovane primjere u datoteku `primjeri.txt` tako da u svakom retku budu vrijednosti x, y te indeks izvora (broj od 1 do 5).

(b) Implementirajte algoritam k-srednjih vrijednosti (pritom nemojte koristiti gotovu implementaciju) i pokrenite ga nad vašim skupom uz $K = 5$. Kako ste odabrali početna središta? Je li vaš algoritam konvergirao? Ako jest, u koliko je iteracija konvergirao? U izvještaju napišite ostvarenu vrijednost kriterijske funkcije te priložite sliku grupiranih primjera (tako da primjeri iz iste grupe budu označeni na identičan način).

(c) U ovom slučaju podatci su umjetno generirani i znamo da je pravi broj grupa 5. U općenitom slučaju broj grupa je nepoznat i potrebno je primijeniti neki od načina provjere grupiranja. Izračunajte optimalnu vrijednost broja grupa K^* za algoritam k-srednjih vrijednosti primjenom Akaikeovog informacijskog kriterija. Ispitajte vrijednosti za K u intervalu iz skupa $\{1, 10\}$. U izvještaju napišite izračunate vrijednosti Akaikeovog kriterija.

(d) Korištenjem algoritma maksimizacije očekivanja pokušajte odrediti parametre Gaussove mješavine uz $K = 5$. Koliko je iteracija potrebno da bi algoritam konvergirao? Kolika iznosi log-izglednost nakon završetka izvođenja algoritma?

(e) Vrednovanje rezultata grupiranja općenito nije jednostavan problem. No u ovom slučaju, budući da su primjeri umjetno generirani i da za svaki primjer znamo koja ga je komponenta generirala, zapravo raspolažemo označenim skupom primjera, pa vrednovanje možemo napraviti pomoću Randovog indeksa. Randov indeks R definiran je kao omjer broja parova primjera koji su nakon grupiranja završili u ispravnim grupama i ukupnog broja parova primjera:

$$R = \frac{a + b}{\binom{N}{2}}.$$

Vrijednost a je broj parova primjera koji imaju istu oznaku i koji su nakon grupiranja završili u istoj grupi. Vrijednost b je broj parova primjera koji imaju različite oznake, a koji su nakon grupiranja završili u različitim grupama. N je broj primjera, a $\binom{N}{2}$ ukupan broj mogućih parova.

Izračunajte Randov indeks za rezultat grupiranja dobiven algoritmom k-srednjih vrijednosti i algoritmom maksimizacije očekivanja. Uzevši u obzir da je odabir početnih središta nedeterminističan, oba algoritma pokrenite po 10 puta i zabilježite najbolju vrijednost Randovog indeksa za oba algoritma. Komentirajte rezultat.