3. domača naloga pri predmetu ITAP

Anamarija Potokar

24. junij 2025

1. naloga

a)

Celovita razdalja je definirana kot največja možna razdalja med katerimakoli dvema elementoma iz dveh gruč. Celovita razdalja spodbuja kompaktne, zaokrožene gruče, združevanje poteka počasneje (kot npr. pri uporabi enojne razdalje), saj se združijo šele tiste gruče, katerih vsi pari so si dovolj blizu. Opisane lastnosti držijo za levi dendrogram, zato bi zanj rekla, da pripada hierarhičnemu razvrščanju s celovito razdaljo.

b)

Na desnem grafu vidimo, da sta gruči jasno ločeni in bolj kompaktni, takšne rezultate pa pričakujemo pri celoviti razdalji, kjer se gruče združujejo le, če so tudi njihove najbolj oddaljene točke dovolj blizu, to pa preprečuje nastanek velikih, razpotegnjenih gruč. Za levi graf pa bi lahko sklepali, da je bila uporabljena enojna razdalja, saj je skoraj vse točke združila v eno gručo, samo ena točka je ostala ločena. To se namreč pogosto zgodi pri razvrščanju z enojno razdaljo, kjer se točke verižijo, ker se združujejo glede na najbližjo razdaljo, ne glede na to, kako razpršene so ostale točke v gruči.

c)

Na levem grafu je vijolična točka jasno blizu skupine rdečih točk, ki pa so izolirane od ostalih skupin. Te rdeče točke tvorijo eno gosto gručo, skupaj z vijolično točko. Ta skupina torej ustreza zeleno obarvanemu delu dendrograma in se pozno priključi ostalim skupinam (ker je daleč stran, kar je tudi značilnost razvrščanja z enojno razdaljo). Znotraj te zelene gruče pa imamo levi del, kjer se 6 točk združi hitro, kar ustreza grafičnemu prikazu na levi, in ostaneta le še 2 točki z indeksoma 4 in 28, ki sta pridruženi postopoma. Indeks 4 se prej priključi gruči rdečih točk, torej bo on ustrezal vijolični točki iz levega grafa.

d)

ENOJNA RAZDALJA: rdeča in bež se združita, ko bo najbližji par točk med rdečo in bež bližje kot katerikoli drugi par med še neločenimi gručami. Na sliki vidimo, da je ena rdeča točka (skrajna desna rdeča točka) zelo blizu leve bež točke, razdalja izgleda približno 1.5 enote. To pomeni, da se bosta gruči združili, ko bo razdalja med njima postala najmanjša med vsemi pari gruč, torej približno pri razdalji 1.5 enot.

CELOVITA RAZDALJA: upošteva se največja razdalja med poljubnimi pari točk iz obeh gruč. Do združitve zdaj pride pri večji razdalji, in sicer pri približno 9 enotah.

e)

Če najprej pogledamo levi dendrogram, sta najdaljši veji na razdalji med 4 in 8, in če dendrogram tam razrežemo, dobimo dve gruči. Pri srednjem dendrogramu lahko režemo na razdalji med 15 in 20, kar nam da dve gruči, ali na razdalji od 10 naprej, kar nam pa da tri gruče. Vendar pa se za drugo možnost pri srednjem dendrogramu ne bomo odločili, saj so potem podatki že manj raznoliki, kot so bili pri dveh gručah. Pri desnem dendrogramu podobno kot pri levem režemo na razdalji nad 8 in spet dobimo dve gruči. Na podlagi teh treh dendrogramov bi rekla, da imamo dve gruči.

2. naloga

a)

Definirala sem ustrezno funkcijo, ki vrne slovar povprečnih vrednosti za vsak razred in povprečno vrednost vseh podatkov.

b)

Za dane podatke sem uporabila prej napisano funkcijo, da sem izračunala povprečje vsake števke, ter ta povprečja tudi vizualizirala. Povprečja so bila skladna z dejanskimi števkami.

c)

Napisala sem funkciji, ki izračunata matriko razpršenosti znotraj razredov S_W in matriko razpršenosti oz. razlik med razredi S_B.

d)

Dokončala sem definicijo za razred LDA.

e)

Ker je LDA omejen s $(k-1)$ komponentami za k razredov, sem uporabila 9 komponent, ki je največje smiselno število komponent, da sem ohranila vso možno ločljivostno informacijo. Transformirala sem testne podatke v LDA prostor in nato vizualizirala prvi dve komponenti. Videti je bilo, da imamo več različnih razredov zmešanih na istem območju, torej med njimi ne obstajajo tako jasne meje, vendar pa se to zdi smiselno, saj prihaja do prekrivanja npr. števk 7 in $9, 3, 5$ in $8, ...$, ki so si vizualno podobne.

f)

Po preverjanju točnosti sem dobila sledeče rezultate:
Točnost na originalnih podatkih: 0.9040

Točnost na LDA podatkih: 0.8545
Točnost na PCA podatkih: 0.7894

Rezultati so kot po pričakovanjih:

- originalni podatki: pričakovana je najvišja točnost, ker imamo vseh 784 spremenljivk (torej celotno sliko)
- LDA: točnost je nekoliko nižja, a še vedno zelo visoka, glede na to, da uporabljamo le 9 dimenzij. LDA optimizira projekcijo tako, da maksimizira ločljivost med razredi, kar je zelo učinkovito za klasifikacijo. Majhen padec točnosti je logičen, saj izgubimo nekaj podrobnosti pri projekciji, vendar pa ohranimo bistveno strukturo med razredi. Še vseeno pa je rezultat super.
- PCA: ohranja dimenzije z največ variance, vendar ne ve, kaj so razredi, torej: najbolj spremenljivi piksli /= najbolj ločujoči piksli med števkami; zato je naravno, da je rezultat slabši kot pri LDA.

Vidimo prednost LDA pred PCA za klasifikacijsko nalogo. Originalni podatki so seveda najbolj informativni, ampak so dimenzijsko zahtevni, LDA pa ponuja dobro ravnovesje med kompaktnostjo in točnostjo.

g)

Kot pričakovano so LDA rekonstruirane slike zelo slabe, saj LDA ni zasnovan za rekonstrukcijo. LDA se namreč trudi ohraniti tiste informacije, ki ločujejo razrede, ostalo pa zanemari, saj za klasifikacijo ni pomembno. Torej ko rekonstruiramo, dobimo nazaj samo tisto, kar je pomagalo ločiti razrede, ne pa celotne slike.

PCA rekonstruirane slike so boljše, saj PCA ohranja največ variance = razlike med vzorci, torej ohrani splošno strukturo pikslov. Z ohranjanjem največjih komponent variance dobimo ohranjene oblike števk. Vendar pa tudi PCA rekonstruirane slike niso prav dobre, saj s tem, ko uporabimo samo 9 komponent od 784, izgubimo veliko podrobnosti, robovi postanejo zamegljeni, nekatere črte manjkajo...

h)

PCA bi uporabili, kadar je glavni cilj ohraniti čim več informacij (variance) iz podatkov brez upoštevanja razredov. PCA torej uporabimo, kadar:

- nimamo ciljne spremenljivke (nenadzorovano učenje, npr. vizualizacija, gručenje)
- želimo vizualizirati podatke, ne da bi vedeli, kaj je razred
- želimo zmanjšati dimenzijo pred uporabo modela, ki ni občutljiv na razredno strukturo
- delamo s podatki, kjer je vsaka dimenzija šum in želimo ohraniti le "močne vzorce".

LDA bi uporabili, ko želimo maksimalno ločiti razrede v podatkih (nadzorovana metoda). LDA torej uporabimo, kadar:

- imamo podatke z oznakami in želimo napovedati, v katerem razredu bodo točke
- želimo izboljšati klasifikacijo
- potrebujemo učinkovitejšo vizualizacijo klasifikacijskih razlik.

1 3. naloga

V datoteki dn3.csv je podana podatkovna množica študentov z različnimi osebnimi, akademskimi in socio-ekonomskimi značilnostmi. Ciljna spremenljivka (Target) opisuje stanje posameznika: ali je študij zaključil (Graduate), ga še vedno opravlja (Enrolled) ali pa je študij prekinil (Dropout). Naloga je bila izdelati napovedni model, ki na podlagi danih značilnosti čim bolj točno napove ciljno kategorijo. V okviru naloge smo izvedli več faz modeliranja, testirali različne algoritme in analizirali učinek predobdelav in izboljšav.

Najprej sem podatke naložila upoštevajoč ločilo ; in odstranila vrstice z manjkajočimi vrednostmi. Kategorične spremenljivke sem zakodirala z LabelEncoder-jem in podatke razdelila na učno (80%) in testno (20%). Pri tem sem uporabila stratify=y, da se je ohranilo razmerje razredov.

Nato sem najprej natrenirala osnovne modele logistične regresije, odločitveno drevo in naključni gozd. Pri tem sem analizirala napako MSE, accuracy, F1-score in matrike zmede.

Nato sem poskusila s prvo izboljšavo, ki je bila skaliranje podatkov, in razlike med skaliranimi in osnovnimi modeli izpisala v obliki tabele. Kjer je bila razlika v točnosti pozitivna, je pomenilo, da skaliranje izboljša model, saj je točnost skaliranega modela večja. Kjer je bila razlika v napaki MSE pozitivna, je pomenilo, da skaliranje poslabša model, saj je napaka skaliranega modela večja. Kjer je bila razlika v F1 (macro/weighted) pozitivna, pa je spet pomenilo, da skaliranje izboljša model.

Za naslednjo izboljšavo sem želela odstraniti šlabe atribute", npr. tiste, ki imajo preveč manjkajočih vrednosti, saj lahko odstranitev izkrivlja rezultate modela in povzroči, da model ignorira pomembne vzorce. Slab atribut tudi pomeni, da ima zelo nizko variabilnost, namreč če je atribut skoraj konstanten, potem se model iz tega ne mora naučiti nič koristnega. Preverila sem torej stolpce z veliko manjkajočimi vrednostmi ($> 10\%$) in nizko varianco (< 0.01). Ker noben stolpec ni ustrezal tem kriterijem, nisem odstranila nobene spremenljivke.

V naslednjem koraku sem se lotila uravnoteženja razredov z uporabo SMOTE: z metodo SMOTE sem generirala umetne primere za premalo zastopane razrede. Izdelala sem tudi tabelo razlik (SMOTE - original).

Nazadnje sem uporabila še model XGBoost in primerjala rezultate originalnih modelov z modelom XGBoost.