Poslovna inteligenca

3. izpitni rok

17. september 2013

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	6	Vsota
Vrednost	9	3	9	5	5	6	37
Točk							

1. Gimnazijski sošolci se odločajo za izbor gostilne za obletnico mature. Samo štirje so oddali svoje glasove:

   Gostilne

	Gostilne			
	Pr'Metki	Pr'Janezu	Pr'Lojzki	Pr'Petru
Žan	3	5	4	8
Zarja	2	4	6	3
Taras	5	5	6	4
Eva	4	4	1	4

   Vsako od predlaganih gostiln so ocenjevali na lestvici od 0 do 10, kjer je 10 najboljša ocena.

[3] (a) Poiščite pareto-optimalne in sub-optimalne gostilne.

[2] (b) Katero gostilno bi izbrali, če bi se odločalo po metodi Harsany-ja?

[2] (c) Katera gostilna bi bila najbolj primerna po metodi Nash-a?

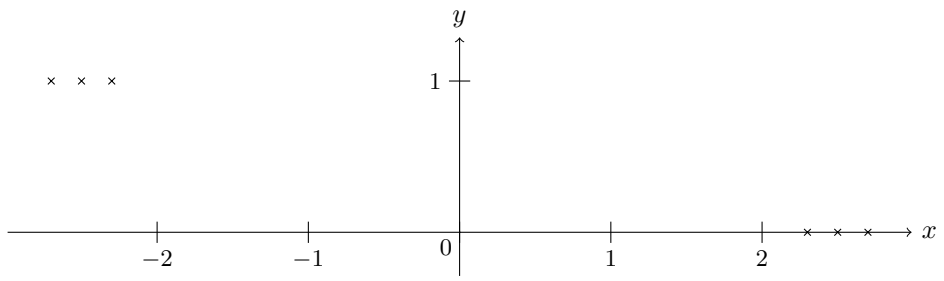
[2] (d) Če bi Zarja, kot glavna organizatorka, imela dvakratno utež kot ostali (ostali pa imajo enake uteži), katera gostilna bi bila zmagovalna?

[3] 2. (a) Na voljo imate podatkovni nabor z $m = 1.000.000$ primeri in $n = 200.000$ atributi ter zveznim razredom. Za gradnjo napovednega modela bi želeli uporabiti linearno regresijo ter oceniti vrednosti parametrov Θ. Kateri postopek ocene parametrov boste uporabili, analitični pristop z normalno enačbo ali iterativni stohastični pristop z gradientnim sestopom? Zakaj?

[3] 3. (a) Označi, ali so sledeče izjave glede logistične regresije resnične ali ne.

- Z regularizacijo ne moremo poslabšati rezultatov na učni množici.
- Z regularizacijo ne moremo poslabšati rezultatov na testni množici.
- Z dodajanjem novih atributov v model (npr. zmnožkov obstoječih atributov) preprečimo pretirano prilagajanje podatkov učni množici.

[3] (b) Imamo podatke s šestimi primeri in enim atributom (x, glej skico). Napovedati želimo razred y. Dvakrat uporabimo logistično regresijo: prvič z zelo majhno vrednostjo regularizacijskega koeficienta λ, drugič z zelo veliko. V koordinatni sistem vrišite krivulji, ki opisujeta napovedi logistične regresije $P(Y = 1)$ pri veliki in majhni vrednosti λ.



[3] (c) Osnovno tehniko logistične regresije uporabljamo na dvorazrednih podatkih. Kako prilagodimo postopek za podatke, kjer je razredov več (npr. pet)?

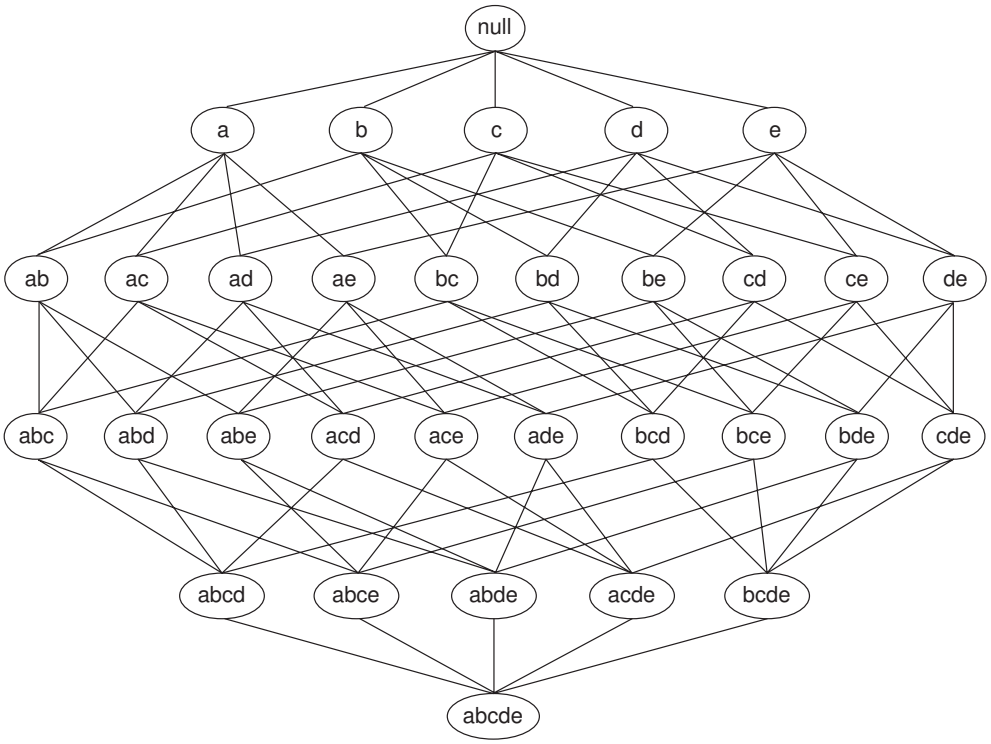
[5] 4. Datoteke (vse s končnicami .txt), ki dejansko vsebujejo programsko kodo, želimo razvrstiti glede na programski jezik: Python ali C. Ko smo že opravili polovico dela tako, da smo polovici datotek priredili ustrezne končnice (.py ali .c), smo z dolgotrajnim ročnim pregledovanjem datotek obupali. Ročno se nam datotek ne da več pregledovati! Namesto tega bi želeli ustrezni program za avtomatsko razvrščanje, ki za razvrstitev še neuvrščenih datotek lahko uporabi vedenje o že razvrščeni polovici. Kako bi lahko zasnovali program, s katerim bi ugotovili, ali je neka koda v Pythonu ali C-ju? Vsebine datoteke pri tem ne bomo ročno pogledali (in je ne poskušamo prevajati ali poganjati)? Ali lahko nalogo rešiš z uporabo programa za stiskanje (kompresijo) besedil? Kako?

[5] 5. Dani so transakcijski podatki v obliki nakupovalnih košaric:

ID	kupljeni izdelki
1	$\{a, b, c, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

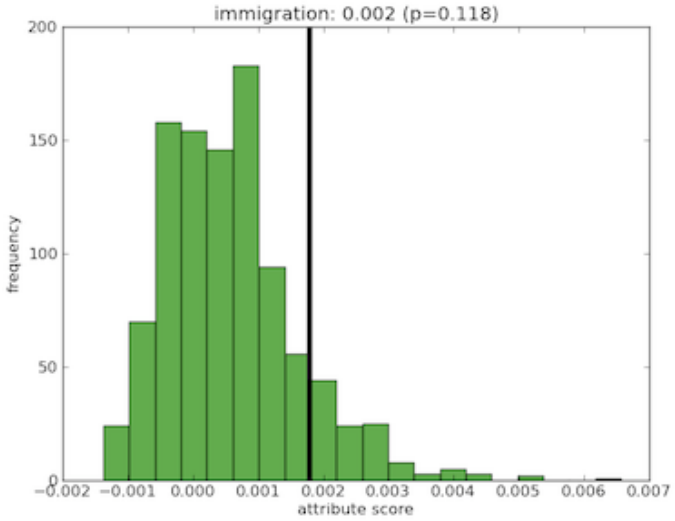
Algoritem *Apriori* gradi skupine izdelkov velikosti $k + 1$ z združevanjem pogostih skupin (frequent itemsets) velikosti k. Za algoritem *Apriori* s spodnjo mejo podpore (minimum support) 20% na skici označi skupine izdelkov s črko:

- N, če *Apriori* za to skupino izdelkov sploh ne računa podpore.
- F, če *Apriori* računa podporo za to skupino izdelkov in ugotovi, da je večja ali enaka minimalni.
- I, če *Apriori* računa podpora za to skupino izdelkov in ugotovi, da je manjša od minimalne.



$$\sigma(X) = |\{t_i | X \subseteq t_i,\ t_i \in T\}| \qquad s(X \to Y) = \sigma(X \cup Y)/N \qquad c(X \to Y) = \sigma(X \cup Y)/\sigma(X)$$

6. Koristnost atributov za neki klasifikacijski problem smo ocenjevali tako, da smo za vsak posamezen atribut s permutacijskim testom pridobili referenčno (ničelno) porazdelitev ocene koristnosti tega atributa. Primer take porazdelitve za izbrani atribut "immigration"kaže spodnji histogram. Uporabljena ocena atributa ("atributa score") je taka, da je atribut pri njenih višjih vrednostih bolj informativen. V spodnjo porazdelitev smo vrisali tudi oceno informativnosti (attribute score) tega atributa na originalnih (nepermutiranih) podatkih. Njegova ocena je znašala 0.002.



[2] (a) Kaj smo pravzaprav naredili s permutacijskim testom oziroma kako smo pridobili podatke za izris ničelne porazdelitve ocene atributa?

[1] (b) Oceni, kolikokrat smo morali ponoviti permutacijski poskus, da smo dobili podatke za prikazani histogram. Z drugimi besedami, kolikokrat smo morali v ta namen izmeriti informativnost atributa "immigration".

[2] (c) Kaj na sliki pomeni p=0.118?

[1] (d) Bi bilo ta atribut smiselno obdržati v učni množici? Zakaj?

Stran je prazna, da lahko nanjo rešujete nalogo.