Strojno učenje – domaća zadaća 4

UNIZG FER, ak. god. 2012/13.

Zadano: 6.1.2013. Rok predaje: 20.1.2013. do 23.59 sati.

Zadatak 1: Stabla odluke

(a) Analizirajte algoritam ID3 u smislu osnovne tri komponente (model, funkcija gubitka, optimizacijski postupak).

(b) Uporabom algoritma ID3 izgradite stablo odluke za troklasnu klasifikaciju primjera prema kriteriju *"Dobro i isplativo ljetovanje na Jadranu"*. Oznaka primjera $y^{(i)}$ može poprimiti jednu od tri vrijednosti: *da*, *ne*, *možda*. Skup primjera za učenje je sljedeći:

i	Mjesto	Otok	Smještaj	Noćenja	Društvo	Prijevoz	Cijena (kn)	$y^{(i)}$
1	Istra	da	privatni	5	obitelj	auto	4.000	da
2	Kvarner	ne	hotel	14	par	auto	5.000	ne
3	Dalmacija	da	hotel	3	par	avion	8.000	da
4	Dalmacija	ne	privatni	7	par	avion	5.000	ne
5	Istra	ne	privatni	5	obitelj	auto	6.000	da
6	Kvarner	ne	vlastiti	14	ekipa	autobus	3.000	ne
7	Kvarner	da	kamp	7	ekipa	autobus	2.500	možda
8	Dalmacija	da	hotel	7	ekipa	auto	2.000	da
9	Istra	ne	privatni	5	obitelj	auto	4.500	da
10	Kvarner	ne	hotel	14	ekipa	auto	2.500	ne
11	Dalmacija	ne	kamp	21	par	autobus	9.000	možda
12	Istra	ne	privatni	3	obitelj	avion	9.000	da
13	Kvarner	da	hotel	14	par	auto	4.000	da
14	Dalmacija	ne	vlastiti	14	par	auto	3.500	ne
15	Istra	ne	vlastiti	5	par	autobus	2.000	možda

Napomena: U prvom koraku napišite postupak izračuna; u daljnjim koracima izračun ne treba pisati.

(c) Pretpostavite da kod 6. primjera nedostaje vrijednost značajke *Smještaj*. Kako biste riješili taj problem?

(d) Je li algoritam ID3 sklon prenaučenosti? Zašto? Objasnite kako biste riješili problem prenaučenosti kod algoritma ID3.

Zadatak 2: Algoritam k-NN

(a) Klasifikator temeljen na algoritmu 4-nn s težinskim faktorima učen je na sljedećim primjerima iz $\mathbb{R}^3 \times \{0, 1\}$:

$$\mathcal{D} = \big\{\big((4,4,0),1\big),\ \big((4,3,1),1\big),\ \big((6,0,2),1\big),\ \big((5,2,2),0\big),\ \big((5,1,1),0\big),\ \big((7,2,0),0\big)\big\}.$$

Odredite klasifikaciju primjera $\mathbf{x}^{(1)} = (4, 2, 1)$ i $\mathbf{x}^{(2)} = (0, 3, 3)$. Za klasifikaciju se koristi euklidska udaljenost s težinskim faktorom obrnuto proporcionalnim kvadratu udaljenosti između primjera.

(b) Skicirajte (za općenit slučaj) pogrešku učenja i pogrešku generalizacije kao funkciju od k.

(c) Skicirajte (za općenit slučaj) pogrešku učenja i pogrešku generalizacije kao funkciju broja primjera N za $k = 1$ i $k = 3$ (nacrtajte na dva zasebna grafikona).

Zadatak 3: Vrednovanje klasifikatora u Rapid Mineru

U ovom zadatku koristit ćete alat Rapid Miner kako biste izgradili i eksperimentalno vrednovali više klasifikatora za dva različita klasifikacijska problema. Preuzmite i instalirajte Rapid Miner s adrese `www.rapidminer.com` te pročitajte upute. Zatim preuzmite sljedeće skupove podataka:

- *Magic* – na temelju niza numeričkih značajki koje opisuju sliku sa teleskopa MAGIC, potrebno je odrediti radi li se o signalu (klasa g) ili pozadinskom zračenju (klasa h). Koristite datoteku `magic04.csv`. Oznaka klase nalazi se u zadnjem stupcu datoteke. Smanjite skup tako da izdvojite proizvoljno odabranih 3000 elemenata klase g i 2000 elemenata klase h (ovo možete napraviti i u Rapid Mineru, npr. pomoću blokova *Filter Examples* i *Union*). U izvještaju napišite koje ste primjere izdvojili (može biti u obliku raspona npr. 2000–5000 i 14000–16000 iz izvornog skupa).
- *Promoters* – za niz DNK-baza treba odrediti radi li se o karakterističnim djelovima DNK (promotorima) od kojih kreće sinteza proteina. Koristite datoteku `promoters.csv`. Uzorak je prozor od ukupno 57 baza koje mogu poprimati diskretne vrijednosti (C,G,A ili T). Uzorke treba klasificirati u jednu od dvije moguće klase: kao pozitivne ili negativne . Prvi red ulazne datoteke sadrži imena stupaca. Prvi stupac je oznaka klase.

Učitavanje skupova podataka u Rapid Miner obavlja se na način opisan u uputama.

U ovom zadatku bit će potrebno provesti odabir modela (optimizaciju hiperparametara) u kombinaciji s vanjskom unakrsnom provjerom (metodom izdvajanja). U glavnom izborniku odaberite *Open Templates*, otvorite predložak *Optimize Parameters* i proučite kako radi. Proučite na koji način biste promijenili koji se parametri variraju ili raspon u kojemu se oni variraju. Na izlazu *mod* bloka *Optimize Parameters* može se dobiti model učen na cijelom ulaznom skupu uz korištenje optimalnih parametara.

(a) Krenite od predloška *Optimize Parameters*. Izmijenite ga tako da blok *Optimize Parameters* omotate u blok *Split Validation* (za ispitivanje izdvojite 30% primjera uz stratificirano uzorkovanje). Time ste osigurali da se optimizacija parametara obavlja samo na skupu za učenje, dok se ispitivanje obavlja na izdvojenom skupu. Broj preklopa u bloku *X-Validation* unutar bloka *Optimize Parameters* neka je 5.

(b) Koristeći proces koji ste napravili u prethodnom zadatku ispitajte rad sljedećih pet klasifikatora na oba skupa podataka:

- Logistička regresija (*Logistic Regression*) – varirajte hiperparametar C u rasponu od 0.01 do 100 po logaritamskoj skali u 5 koraka; *Napomena:* Isključite opciju *scale*.

- Stroj s potpornim vektorima (*Support Vector Machine*) s polinomijalnom jezgrenom funkcijom drugog stupnja – varirajte parametar C u rasponu od 0.01 do 100 po logaritamskoj skali u 5 koraka; *Napomena:* Isključite opciju *scale*.
- Naivan Bayesov klasifikator (*Naive Bayes*) – varirajte hiperparametar Laplaceovog zaglađivanja (*true* ili *false*). *Napomena:* ako su atributi numerički a niste napravili diskretizaciju, ovaj blok će ju interno sam provesti;
- Algoritam k-najbližih susjeda (*k-NN*) – varirajte hiperparametar k u rasponu od 1 do 150 u 20 koraka;
- Stablo odluke, varijanta C4.5 (*Decision Tree*) – varirajte *Criterion* (*Information gain*, *Gain ratio*), *Minimal leaf size* (u rasponu $\{1, \ldots, 10\}$).

Prva dva klasifikatora ne mogu raditi s diskretnim značajkama, pa je takve značajke potrebno pretvoriti u numeričke. Za skup podataka *Magic* preporučljivo je provesti normalizaciju vrijednosti značajki uporabom bloka *Normalize*. Sve pretvorbe i predobrade podataka najbolje je staviti odmah nakon bloka za učitavanje skupa podataka (a prije bloka *Split Validation*).

Za svaki klasifikator zabilježite dobivenu točnost (*accuracy*) i matricu zabune.

Napomena: Izvještaju treba priložiti datoteke s modelima u Rapid Mineru (po jednu datoteku za svaki par *model – skup podataka*).

(c) U izvještaju napravite dvije tablice, po jednu za svaki skup podataka. U retcima tablice navedite različite klasifikatore, a u stupcima mjere pogreške: točnost, mikro-F1, makro-preciznost, makro-odziv i makro-F1. Mjeru točnosti Rapid Miner izračunava automatski, a ostale mjere trebate izračunati sami na temelju matrice zabune koju Radid Miner daje kao rezultat vrednovanja.

(d) Komentirajte dobivene rezultate: koji je klasifikator najbolji za prvi, a koji za drugi problem? Koja od upotrijebljenih mjera smatrate da najrealističnije ocjenjuje pogrešku klasifikatora? Postoji li razlika između vrijednosti mjera mikro-F1 i makro-F1 te zašto?

(e) Često je korisno provesti postupak odabira podskupa značajki (engl. *feature subset selection*), odnosno od mnogo značajki odabrati one koje su najpogodnije za klasifikaciju. U ovom podzadatku koristiti ćemo postupak odabira filtrom. Krenite od praznog projekta, učitajte skup podataka *Promoters*, pretvorite polinominalne atribute u binominalne (*Nominal to Binominal*), zatim izračunajte težinu svakog atributa kao korelaciju s ciljnim atributom (*Weight by Correlation*). Konačno, izbacite sve osim k atributa s najvećim težinama (*Select by Weights*). Na tako preinačenom skupu provedite procjenu učinkovitosti Bayesovog klasifikatora (s Laplaceovim zaglađivanjem). Za procjenu koristite unakrsnu provjeru s deset preklopa. Eksperimentirajte kako veličina k utječe na učinkovitost klasifikacije.

Odgovorite na pitanja:

(i) Uz koji k dobivate najbolje rezultate? Kolika je razlika u točnosti klasifikatora u odnosu na slučaj gdje se koriste sve značajke?

(ii) Koliko najviše možete smanjiti k a da se točnost ne pokvari značajno (više od 1%) u odnosu na slučaj gdje se koriste sve značajke?

(iii) Ako uvažimo činjenicu da izbor značajki također utječe na točnost modela, je li ovakva procjena točnosti pesimistična ili optimistična? Obrazložite zašto.

(iv) Opišite (riječima; nije potrebno crtati model) što bi trebalo promjeniti u načinu vrednovanja da procjena bude realističnija.

Napomena: Izvještaju priložite datoteku s modelom u Rapid Mineru.