Strojno učenje – domaća zadaća 4

UNIZG FER, ak. god. 2011/12.

Zadano: 5.1.2012. Rok predaje: 17.1.2012. do 17.00 sati.

Zadatak 1: Stabla odluke

(a) Uporabom algoritma ID3 izgradite stablo odluke za klasifikaciju primjera u klasu "Popularan izborni predmet". Primjeri za učenje su sljedeći:

i	Ispit	Projekt	Predavanja	Labosi	Prosjek	$h(\mathbf{x}^{(i)})$
1	pismeni	da	obavezna	ne	visok	1
2	usmeni	da	neobavezna	ne	srednji	0
3	oboje	opcionalno	obavezna	da	nizak	0
4	oboje	ne	neobavezna	ne	visok	1
5	pismeni	ne	obavezna	da	srednji	0
6	usmeni	opcionalno	neobavezna	ne	srednji	1
7	pismeni	ne	obavezna	da	nizak	0
8	pismeni	ne	neobavezna	ne	visok	1

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

(c) Može li se dogoditi da jedan se jedan te isti primjer $\mathbf{x}^{(i)}$ u skupu za učenje pojavi dva puta i to s različitom oznakom $y^{(i)}$? Zbog čega bi se to moglo dogoditi? Kako se taj problem rješava kod stabla odluke? Ilustrirajte na gornjem skupu podataka.

(d) Stablo odluke je neparametarski model. Što to konkretno znači?

(e) Razmatramo skupove primjera za učenje $\mathcal{D} \subset \mathbb{R}^2 \times \{0, 1\}$ za koje *ne* postoji hipoteza

$$h_1(x_1, x_2|\theta_0, \theta_1, \theta_2) = \mathbf{1}\{\theta_1 x_1 + \theta_2 x_2 + \theta_0 \geqslant 0\}$$

koja bi bila konzistentna s primjerima za učenje. Možemo li algoritmom ID3 za svaki takav skup $\mathcal{D}$ naučiti hipotezu h_2 koja jest konzistentna s primjerima za učenje? Obrazložite odgovor.

(f) Odgovorite na gornje pitanje uz pretpostavku da je stablo odluke ograničeno na dubinu 3. Ima li općenito smisla ograničiti stablo na neku dubinu? Kako biste odabrali na koju dubinu ga ograničiti? Kako biste to napravili ako na raspolaganju imate samo navedenih osam primjera za učenje?

Zadatak 2: Algoritam k-NN

(a) Raspolažemo skupom primjera za učenje $\mathcal{D} = \mathbb{R}^2 \times \{0, 1\}$:

$$\mathcal{D} = \big\{((-1, 1), 0), ((-1, -1), 0), ((0, 0), 1), ((1, 1), 1), ((1, -1), 1)\big\}\,.$$

Skicirajte primjere u ulaznom prostoru te granicu između klasa za klasifikatore 1-NN i 3-NN.

(b) Algoritam k-najbližih susjeda može se upotrijebiti i za klasifikaciju multinomijalnih varijabli, ako se upotrijebi prikladna mjera udaljenosti. Jedna mogućnost jest primjere $\mathbf{x}^{(i)}$ kodirati kao binarne vektore, a zatim upotrijebiti euklidsku ili Jaccardovu mjeru udaljenosti. Jaccardova mjera udaljenosti između vektora $\mathbf{x}^{(i)}$ i $\mathbf{x}^{(j)}$ definirana je kao

$$J'(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = 1 - J(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = 1 - \frac{\sum_{k=1}^{n} \mathbf{1}\{x_k^{(i)} \wedge x_k^{(j)}\}}{\sum_{k=1}^{n} \mathbf{1}\{x_k^{(i)} \vee x_k^{(j)}\}}\,.$$

Koristeći primjere iz zadatka 1, odredite klasifikaciju 8. primjera pomoću klasifikatora 3-NN naučenog na prvih sedam primjera uz uporabu (1) euklidske mjere udaljenosti i (2) Jaccardove mjere udaljenosti.

(c) Binarna vektorizacija multinomijalne varijable prikladna je kada sve razlike u vrijednostima multinomijalne varijable želimo tretirati jednako. Naime, ako se dva primjera razlikuju u vrijednosti samo jedne multinomijalne značajke, njihova će udaljenost biti jednaka neovisno o vrijednostima te značajke. Ako je svaki primjer sastavljen od n multinomijalnih značajki, koliko će iznositi ta udaljenost u slučaju euklidske udaljenosti, a koliko u slučaju Jaccarove udaljenosti?

U nekim situacijama međutim ne želimo razlike u vrijednostima tretirati jednako i tada treba drugačije definirati mjeru udaljenosti. Je li to slučaj s problemom iz zadatka 1a? Ako smatrate da jest, predložite bolju mjeru udaljenosti.

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:

- *Statlog (Vehicle Silhouettes)* – na temelju niza numeričkih značajki koje opisuju siluetu vozila treba odrediti o kojoj se vrsti vozila radi: *van*, *bus*, *saab* ili *opel*. Skup dolazi u 9 datoteka: proizvoljno odaberite 6 datoteka i spojite ih u jednu datoteku (u izvještaju napišite koje ste datoteke odabrali).
- *Splice* – za niz DNK-baza treba odrediti radi li se o karakterističnim granicama među djelovima koji se odrezuju prije sinteze proteina. Uzorak je prozor od 30 baza lijevo i desno od promatrane baze (ukupno 60 značajki) koje mogu poprimati diskretne vrijednosti. Uzorke treba klasificirati u jednu od tri moguće klase: *IE boundary*, *EI boundary* i *None*. Izvorni skup podataka nije u CSV-datoteci; CSV-datoteku možete preuzeti ovdje. Prvi red ulazne datoteke sadrži imena stupaca. Prvi stupac je oznaka klase, a drugi stupac je identifikator (koji pri učenju modela treba zanemariti).

Učitavanje skupova podataka u Rapid Miner obavlja se na način opisan u uputama. *Napomena:* Kod skupa podataka *Splice*, svi stupci moraju kao tip imati *polynominal* ili *nominal* (ovo treba provjeriti prije uvoza).

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 kojem 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 10.

(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 10 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 10 koraka;
- 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. Također, prva dva klasifikatora ne mogu raditi s više od dvije klase, pa ih je potrebno omotati blokom *Polynominal by Binomial Classification*. Za skup podataka *Statlog* 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 (v. poglavlje 6 u uputama za Rapid Miner). Pronađite optimalan podskup značajki za Bayesov klasifikator (uz Laplaceovo zaglađivanje) na problemu *Splice*. Odabir treba načiniti metodom omotača korištenjem odabira unaprijed (engl. *forward selection*). U tu svrhu iskoristite blok *Wrapper Split Validation*, prema uputama. Iz skupa primjera 30% primjera izdvojite za ispitivanje (koristite stratificirano uzorkovanje). Koja je veličina optimalnog podskupa značajki i kolika je razlika u točnosti klasifikatora u odnosu na slučaj iz podzadatka b u kojem se koriste sve značajke?