

Sklop 5: Validacija modelov strojnega učenja

Uvod

Ko zgradimo klasifikator je ključnega pomena, da validiramo kako le ta deluje. V idealnem primeru, ko imamo na voljo dovolj podatkov le te razdelimo v dve množici – učno in testno. Klasifikator zgradimo s podatki učne množice in ga preizkusimo s podatki testne množice.

Če pa nimamo na voljo dovolj podatkov lahko, za testiranje učinkovitosti klasifikatorja uporabimo metodo navzkrižne validacije (*ang. cross-validation*). V postopku navzkrižne validacije se za učenje klasifikatorja in njegovo validacijo uporabi iste podatke. V začetku postopka določimo število pregibov m danega podatkovnega nabora. Število pregibov pomeni število delitev podatkovnega nabora na podmnožice. V nadaljevanju postopka vzorce v podatkovnem naboru naključno premešamo in razdelimo v m enako velikih podmnožic. Po opravljeni razdelitvi se postopek učenja požene m -krat. Pri vsaki iteraciji se $m-1$ podmnožic podatkovnega nabora uporabi za učenje klasifikatorja, preostanek pa za njegovo testiranje. Rezultat vsakega testiranja se kot število pravilno in napačno razvrščenih vzorcev za določen razred zabeleži v **matriko razvrščanja** (*ang. confusion matrix*). Ob koncu postopka navzkrižne validacije rezultate posameznih testiranj zberemo v matriki. Dobljeni rezultat je dober približek zanesljivosti klasifikatorja. Glede na število delitev dane učne množice poimenujemo metodo validacije m pregibna navzkrižna validacija.

Dejanski Napovedani	setosa	versicolor	virginica	class precision
setosa	50	0	0	100.00%
versicolor	0	47	4	92.16%
virginica	0	3	46	93.88%
class recall	100.00%	94.00%	92.00%	

Tabela: Primer matrike razvrščanja za klasifikator Naive Bayes za podatke Iris.

Matrika razvrščanja vsebuje natančne podatke o delovanju klasifikatorja, vendar je posamezne matrike med sabo težko primerjati. Zato za lažjo primerjavo metod iz matrike izračunamo različne parametre: Natančnost, Občutljivost in F1 Oceno, na podlagi katerih lahko klasifikatorje med sabo lažje primerjamo.

Natančnost - Delež pravilno razvrščenih (*ang. precision*)

$$\text{Natančnost} = \frac{\text{število pravilno razvrščenih}}{\text{število vseh}}$$

Občutljivost - Delež najdenih (*ang. recall*)

$$\text{Občutljivost} = \frac{\text{število pravilno razvrščenih}}{\text{število vseh relevantnih}}$$

F1 Ocena – združena ocena natančnosti in občutljivosti

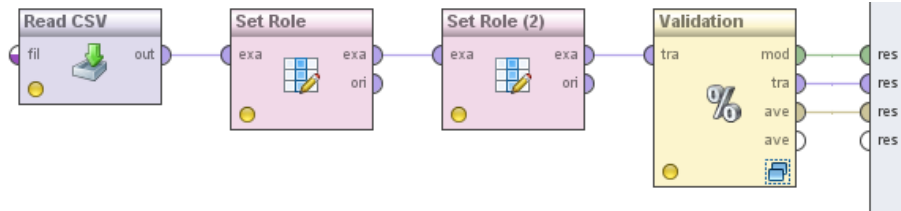
$$F_1 = 2 \cdot \frac{\text{Natančnost} \times \text{Občutljivost}}{\text{Natančnost} + \text{Občutljivost}}$$

Glede na naravo problema se odločimo, kateri performančni parameter bomo pri klasifikatorju opazovali in optimirali.

Naloga

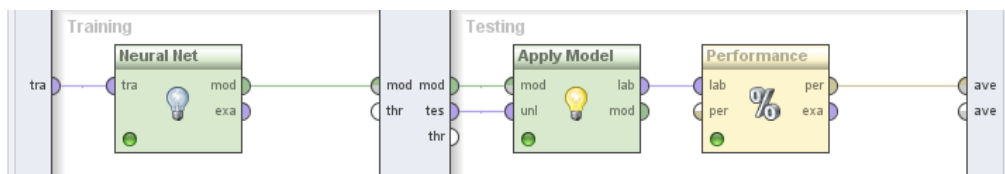
Med sabo primerjajte dva klasifikatorja – Naivni Bayesov klasifikator in Nevronska mrežo za nabor podatkov Atleti.

Naložite podatke in nastavite vlogo **id** značilki Player_name, ter vlogo **label** značilki Team_Value.



Uporabite operator *Validation* in ga zgradite tako, da boste v trening delu uporabljali raziskovane klasifikatorje *Neural Net* in *Naive Bayes*. Izberite 10 pregibno navzkrižno validacijo.

V testnem delu uporabite operator *Apply Model* in izmerite njegovo delovanje z operatorjem *Performace*. Podatke zapišite v tabeli.



Dejanski \ Napoved	Superstar	Contributor	Franchise	Role Player	class precision
Superstar					
Contributor					
Franchise					
Role Player					
class recall					

Tabela: Rezultat validacije klasifikatorja Naive Bayes

Dejanski \ Napoved	Superstar	Contributor	Franchise	Role Player	class precision
Superstar					
Contributor					
Franchise					
Role Player					
class recall					

Tabela: Rezultat validacije klasifikatorja Neural Net

Glede na dejstvo, da igralci veliko stanejo in da želite z večjo gotovostjo napovedati kdo je »Superstar« navedite, za kateri klasifikator bi se odločili. _____

Zakaj? _____

Vir

- Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining. Practical Machine Learning. Tools and Techniques. Third Edition

- <http://rapid-i.com/>

- <https://sites.google.com/site/dataminingforthemasses/>

Vaja pripravljena v okviru KCUI delavnice na UL-FE