|  |
| --- |
|  |
| KOLEGIJ  **UVOD U STROJNO UČENJE** |
| EVALUACIJSKE MJERE ZA OCJENU KVALITETE KLASIFIKACIJSKIH ALGORITAMA |
| Vježba br. 4. |

**BINARNA (DVOKLASNA) KLASIFIKACIJA**

**ZADACI:**

1. Učitati u Weku skup podataka **breast-cancer** i opisati ga:
   1. Izvor skupa podataka (može se vidjeti ako se breast-cancer.arff file otvori npr. u Excelu)

Izvor podataka je Onkološki institut u Ljubljani

1.2. Koliko primjera ima? 286

1.3. Koji su atributi i njihove vrijednosti; opisati značenje atributa i njihovih vrijednosti.

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

3. menopause: lt40, ge40, premeno.

4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,

45-49, 50-54, 55-59.

5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,

27-29, 30-32, 33-35, 36-39.

6. node-caps: yes, no.

7. deg-malig: 1, 2, 3.

8. breast: left, right.

9. breast-quad: left-up, left-low, right-up, right-low, central.

10. irradiat: yes, no

1.4. Ima li atributa s nedostajućim vrijednostima – identificirati ih. Koliko ima primjera s atributima kojima nedostaju vrijednosti? 9 (21, 32, 51, 55, 72, 93, 150, 241, 265)

1.5. Koji je klasni atribut i koje značenje imaju njegove vrijednosti?

Class: no-recurence-events/ recurence-events

1.6. Kakva je distribucija primjera po klasama?

No-reccurence-events: 201 (70%)

Reccurence-events: 85 (30%)

1. Izvršiti klasifikaciju skupa podataka na način da se prvih 80% primjera pridijeli skupu za treniranje, a preostalih 20% skupu za testiranje:

*Test options* → *Percentage split %* : 80;

*More options* → označiti opciju: *Preserve order for % Split*

Klasifikaciju izvršiti sljedećim klasifikatorima:

2a) Naivnim Bayesovim klasifikatorom (opcija: *NaiveBayesSimple*)

2b) Stablom odluke J48 (naprednija verzija ID3 stabla)

1. Za oba klasifikatora na temelju dobivene matrice istinitosti izračunati sljedeće evaluacijske mjere (ručno! – izvještaju priložiti proračune):

* Ukupnu točnost (**T**) klasifikatora
* Odziv (Osjetljivost) (**O**) za obje klase
* Preciznost (**P**) za obje klase
* **F1-mjeru** za obje klase
* Prosječne vrijednosti za **O**, **P** i **F1-mjeru**, otežane faktorima distribucije primjera po klasama (eng. *Weighted Average*)

Napomena za računanje otežanih prosječnih vrijednosti:

Ako skup podataka ima Na primjera klase a i Nb primjera klase b, te su npr. osjetljivosti za klasu a i

b: Oa i Ob, otežana prosječna vrijednost osjetljivosti će biti:

Rezultate provjeriti u Weki i unijeti u Tablicu:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **KLASIFIKATOR:** | |
| **NB** | **Stablo odluke J48** |
| **T** | 70,17% | 71,93% |
| Klasa **a** | **O** | 83% | 97,2% |
| **P** | 73,2% | 70% |
| **F1- mjera** | 77,8 | 81,28% |
| Klasa **b** | **O** | 47,6% | 28,57% |
| **P** | 62,5% | 85,71% |
| **F1- mjera** | 54,1% | 42,84% |
| Otežane srednje vrijednosti | **O** | 71.9% | 71.9% |
| **P** | 71.2% | 75.8% |
| **F1- mjera** | 71.2% | 67.2% |

1. Analizirati podatke iz Tablice.

Pomoć:

S obzirom na ukupnu Točnost koji biste klasifikacijski model odabrali?

J48

Da li je to ispravno rješenje? Naime, na temelju značenja vrijednosti ciljnog atributa

(klase) zaključite kod koje klase greška u klasifikaciji ima teže posljedice po pacijenticu

(tj. ‘više košta’)?

S obzirom na doneseni zaključak nadalje zaključiti koju grešku ćemo željeti minimizirati, odnosno za klasu u kojoj greške nose teže posljedice, za koju mjeru ćemo željeti da ima što veću vrijednost?

S obzirom na donesene zaključke, koji od dva klasifikacijska modela biste odabrali?

Nije ispravno riješenje. Greška false negative za klasu b je puno gora. Biramo klasifikator sa većom osjetljivošću na klasu b, a to je NB

**VIŠEKLASNA KLASIFIKACIJA**

**3 - KLASNA KLASIFIKACIJA**

**ZADACI:**

1. Učitati u Weku skup podataka **iris** i opisati ga:

5.1. Izvor skupa podataka (otvoriti iris.arff file u Excelu)

(a) Creator: R.A. Fisher

(b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)

(c) Date: July, 1988

5.2. Koliko primjera ima? 150

5.3. Koji su atributi i njihove vrijednosti; opisati značenje atributa i njihovih vrijednosti.\*

1. sepal length in cm

2. sepal width in cm

3. petal length in cm

4. petal width in cm

5.4. Ima li atributa s nedostajućim vrijednostima? 0

5.5. Koji je klasni atribut i koje značenje imaju njegove vrijednosti?

Skup podataka predvđa klasu cvijeta perunike na osnovu latica i listiće glave cvijeta.

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

5.6. Kakva je distribucija primjera po klasama?

Jednaka: 50,50,50 (33,3% po klasi)

1. Izvršiti klasifikaciju skupa podataka na način da se prvih 64% primjera pridijeli skupu za treniranje, a preostalih 36% skupu za testiranje:

*Napomena:*

***Ne*** *označiti opciju: More options* → *Preserve order for % Split*

Objasniti razlog zašto nema smisla sačuvati raspored primjera prilikom podjele skupa

podataka na primjere za treniranje i primjere za testiranje!

Klasifikaciju izvršiti Naivnim Bayesovim klasifikatorom (opcija: *NaiveBayesSimple*).

1. Na temelju dobivene matrice istinitosti izračunati sljedeće evaluacijske mjere (ručno! – izvještaju priložiti proračune).

* Ukupnu točnost (**T**) klasifikatora
* Odziv (Osjetljivost) (**O**) za obje klase
* Preciznost (**P**) za obje klase
* **F1-mjeru** za obje klase
* Prosječne vrijednosti za **O**, **P** i **F1-mjeru**, otežane faktorima distribucije primjera po klasama (eng. *Weighted Average*)

Rezultate provjeriti u Weki i unijeti u Tablicu:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **EVALUACIJSKE MJERE:** | | |
| **Odziv (O)** | **Preciznost (P)** | **F1-mjera** |
| Klasa **a** | 100% | 100% | 100% |
| Klasa **b** | 95% | 90.476 | 92.6829 |
| Klasa **c** | 88.888% | 98.076% | 94.11% |
| Otežane  Srednje vrijednosti | 94.444 | 94.5119 | 94.433 |

1. T = 94.444 Da li je u ovom slučaju T adekvatna mjera za ocjenu modela? Obrazložiti odgovor.

Da. Budući da se radi o klasifikaciji cvijeća, gdje su posljedice za FN i FP gotovo jednake, točnost je adekvatna mjera.

**MULTI - KLASNA KLASIFIKACIJA**

**ZADACI:**

1. Iz dosadašnjih primjera vidljivo je da je ‘ručno’ računanje evaluacijskih mjera zamorno (za 2 i 3 klase). A kod multi-klasnih klasifikacija (primjerice, da imamo 10 ili više klasa, što nije rijedak slučaj u praksi), ručno računanje bi bilo i nemoguće.

Stoga je ovaj zadatak sljedeći:

* U nekom programskom jeziku kojeg preferirate (C, Python, Matlab, bilo koji) napisati program koji omogućava korisniku da upiše matricu istinitosti te na temelju nje računa i ispisuje evaluacijske mjere (ukupnu T, te O, S, P i F1-mjeru za svaku klasu zasebno, te srednje vrijednosti otežane distribucijom primjera po klasama)
* Provjera ispravnosti programa:

Skinuti s interneta besplatno dostupan skup podataka **Car** (link: <https://www.cs.ubc.ca/labs/beta/Projects/autoweka/datasets/> ). Proučiti i opisati što predstavlja skup podataka. U Weki klasificirati skup podataka (odabrati po želji klasifikator te klasifikacijski model trenirati na skupu podataka train.arff, a testirati na skupu podataka test.arff). Matricu istinitosti učitati u svoj program i dobiti evaluacijske mjere. Dobivene rezultate usporediti s evaluacijskim mjerama dobivenim u Weki.

* Primjenom programa izračunati i ispisati evaluacijske mjere za 9-klasnu klasifikaciju čija je tablica istinitosti dana na Slajdu br. 83 (Slajdovi iz predavanja).