Skup uzoraka za učenje i skup uzoraka za ispitivanje

Uvod u raspoznavanje uzoraka

Prof. dr. sc. Slobodan Ribarić

Skup uzoraka za učenje i skup uzoraka za ispitivanje

Skup uzoraka za učenje – označeni uzorci, tj. uzorci s poznatom klasifikacijom (uzorci s “labelom”)

Važna pretpostavka – u uzorcima za učenje sadržana je većina informacija o svojstvima razreda kojima uzorci pripadaju.

(Uzorci za učenje trebali bi biti tipični predstavnici razreda kojima pripadaju!)

Skup uzoraka za učenje i skup uzoraka za ispitivanje

- Za svaki od uzoraka u skupu uzoraka za učenje zahtijeva se:

  - dovoljnost informacije o razredu kojem pripada

  - vremenska postojanost

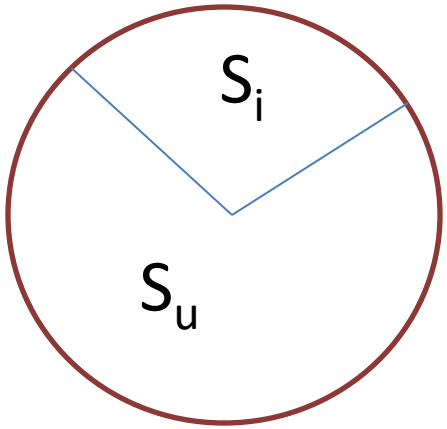
  - geometrijska postojanost (mala udaljenost između uzoraka u prostoru značajki znači i malu razliku u svojstvima objekta koji se razvrstava)

Skup uzoraka za učenje i skup uzoraka za ispitivanje

- Odnos: skup uzoraka za učenje – skup uzoraka za ispitivanje

1. Holdout metoda

Ako imamo dovoljno velik skup uzoraka s poznatom klasifikacijom S možemo napraviti sljedeće:



S_u – skup uzoraka za učenje

S_i – skup uzoraka za ispitivanje

$S = S_u \cup S_i$

$S_u \cap S_i = \emptyset$

Obično:

2/3 S za učenje

1/3 S za ispitivanje

#S = N ukupan broj uzoraka

Skup uzoraka za učenje i skup uzoraka za ispitivanje

- Glavni nedostatak Holdout metode:

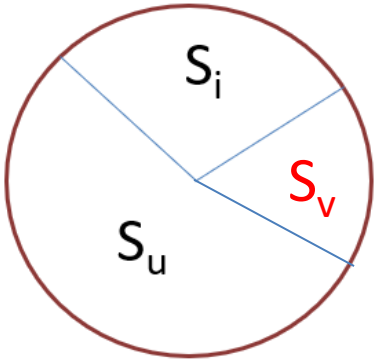
  - smanjeni skupovi podataka za učenje i ispitivanje

  - odluka – koliko od *N* raspoloživih uzoraka dodijeliti skupu uzoraka za učenje a koliko skupu uzoraka za ispitivanje

- Vjerojatnost pogreške klasifikatora koji se oblikuje uporabom konačnog skupa za učenje ($N < \infty$) je uvijek veća negoli je odgovarajuća asimptotska vjerojatnost pogreške ($N \rightarrow \infty$).

Skup uzoraka za učenje i skup uzoraka za ispitivanje

- Vrlo često se pored skupa za učenje i skupa za ispitivanje još koristi i skup za validaciju S_v
- Validacijski skup – skup podataka koji se koristi za „ugađanje” hiperparametara, odnosno arhitekture klasifikatora (npr. ugađanje vrijednosti praga, određivanje broja jedinica u skrivenom sloju NN)
- Za mali skup ispitnih uzoraka ocjena pogreške klasifikatora je nepouzdana



Skup uzoraka za učenje i skup uzoraka za ispitivanje

2. Leave-One-Out metoda

- Metoda pokušava „zaobići” problem podjele skupa označenih uzoraka na skup uzoraka za učenje i ispitivanje
- Učenje se obavlja uporabom *N* - 1 uzoraka, a ispitivanje se obavlja na onom jednom preostalom uzorku!
- Ako je taj uzorak pogrešno klasificiran - inkrementira se brojilo pogrešaka
- Postupak se ponavlja *N* puta ali tako da je svaki put isključen drugi uzorak
- Ukupan broj pogrešaka nas upućuje na procjenu vjerojatnosti pogreške klasifikatora

Skup uzoraka za učenje i skup uzoraka za ispitivanje

- Značajka Leave-One-Out metode

- učenje je na temelju svih uzoraka a istodobno je zadovoljen uvjet održanja nezavisnosti između skupa za učenje i skupa za ispitivanje!

- Analize su pokazale da su rezultati procjene vjerojatnosti pogreške za holdout i leave-one-out metodu vrlo slične za usporedive veličine skupova označenih uzoraka
- Nedostatak metode: velika računska složenost!

Skup uzoraka za učenje i skup uzoraka za ispitivanje

3. Resubstitution metoda (metoda ponovne zamjene)

- isti se skup podataka koristi prvo za učenje a zatim za ispitivanje (!?)

- iz literature: „one no need to go into mathematical details in order to see that procedure is not very fair”

- metoda daje „optimističku” procjenu stvarne vjerojatnosti pogreške

- Za relativno dobru procjenu zahtijeva se dovoljno veliki *N* i dovoljno veliki omjer *N*/n, gdje je n dimenzionalnost prostora značajki

Skup uzoraka za učenje i skup uzoraka za ispitivanje

- Kod vrednovanja klasifikatora često nije dovoljna samo procjena vjerojatnosti pogreške
- Ostale mjere: matrica nedoumice/zbunjenosti (Confusion matrix); osjetljivost (Recall), preciznost (Precision), ukupna točnost (Overall Accuracy)

Pretpostavimo da imamo M > 2 klasifikacijski zadatak

Oblikujemo matricu nedoumice A =[A(i, j)], gdje je element matrice A(i, j) broj uzoraka (vektora) čija je točna oznaka razreda bila i, a on je klasificiran u razred s oznakom j.

Skup uzoraka za učenje i skup uzoraka za ispitivanje

Primjer:
Pretpostavimo da ispitujemo klasifikator za M = 3 razreda. Imamo:
100 ispitnih uzoraka iz razreda ω_1;
120 uzoraka iz razreda ω_2 i
80 ispitnih uzoraka iz razreda ω_3;

- Nakon testiranja klasifikatora dobili smo ovakve rezultate:
  - od 100 uzoraka iz ω_1: 80 ih je ispravno klasificirano u ω_1
    15 u razred ω_2 (pogrešno!)
    5 u razred ω_3 (pogrešno!)
  - od 120 uzoraka iz ω_2: 99 ih je ispravno klasificirano u ω_2
    17 u razred ω_1 (pogrešno!)
    4 u razred ω_3 (pogrešno!)
  - od 80 uzoraka iz ω_3: 60 ih je ispravno klasificirano u ω_3
    7 u razred ω_1 (pogrešno!)
    13 u razred ω_2 (pogrešno!)

Skup uzoraka za učenje i skup uzoraka za ispitivanje

od 100 uzoraka iz ω_1: 80 ih je ispravno klasificirano u ω_1
15 u razred ω_2 (pogrešno!)
5 u razred ω_3 (pogrešno!)
od 120 uzoraka iz ω_2: 99 ih je ispravno klasificirano u ω_2
17 u razred ω_1 (pogrešno!)
4 u razred ω_3 (pogrešno!)
od 80 uzoraka iz ω_3: 60 ih je ispravno klasificirano u ω_3
7 u razred ω_1 (pogrešno!)
13 u razred ω_2 (pogrešno!)

Matrica nedoumice/zbunjenosti:

$$A = \begin{bmatrix} 80 & 15 & 5 \\ 17 & 99 & 4 \\ 7 & 13 & 60 \end{bmatrix}$$

Skup uzoraka za učenje i skup uzoraka za ispitivanje

Na temelju matrice A, izračunavaju se osjetljivost (Recall), preciznost (Precision) i ukupna točnost (Overall Accuracy)

Osjetljivost R_i je postotak uzoraka čija je prava oznaka i (pripadaju razredu ω_i) i bili su ispravno razvrstani u taj razred.

$$\text{Osjetljivost} = \frac{t_p}{t_p + t_n}$$

t_p - istinski pozitivan; t_n - istinski negativan

Skup uzoraka za učenje i skup uzoraka za ispitivanje

$$A = \begin{bmatrix} 80 & 15 & 5 \\ 17 & 99 & 4 \\ 7 & 13 & 60 \end{bmatrix}$$

$$\text{Osjetljivost} = \frac{t_p}{t_p + t_n}$$

$$\text{Osjetljivost}_{\omega_1} = \frac{80}{80 + (15 + 5)} = 0.80$$

$$\text{Osjetljivost}_{\omega_2} = \frac{99}{99 + (17 + 4)} = 0.825$$

$$\text{Osjetljivost}_{\omega_3} = 0.75$$

Skup uzoraka za učenje i skup uzoraka za ispitivanje

$$A = \begin{bmatrix} 80 & 15 & 5 \\ 17 & 99 & 4 \\ 7 & 13 & 60 \end{bmatrix}$$

$$\text{Preciznost} = \frac{t_p}{t_p + f_p}$$

$$\text{Preciznost}_{\omega_1} = \frac{80}{80 + (17 + 7)} = \frac{80}{104} = 0.77$$

$$\text{Preciznost}_{\omega_2} = \frac{99}{99 + (15 + 13)} = \frac{99}{127} = 0.78$$

$$\text{Preciznost}_{\omega_3} = 0.87$$

Skup uzoraka za učenje i skup uzoraka za ispitivanje

Ukupna točnost A_c – postotak podataka koji su bili ispravno klasificirani

$$A_c = \frac{1}{N}\sum_{i=1}^{M} A(i,i)$$

$$A = \begin{bmatrix} 80 & 15 & 5 \\ 17 & 99 & 4 \\ 7 & 13 & 60 \end{bmatrix}$$

$$N = 100 + 120 + 80 = 300$$

$$\mathrm{A}(1,1) = 80;\ \mathrm{A}(2,2) = 99;\ \mathrm{A}(3,3) = 60$$

$$A_c = \frac{1}{N}\sum_{i=1}^{M} A(i,i) = \frac{1}{300}\,(80 + 99 + 60) = 0.796$$

Skup uzoraka za učenje i skup uzoraka za ispitivanje

- Koliko veliki treba biti skup za učenje *N*?

Idealno (teorijski) $N \rightarrow \infty$???

Preporuka za *N*

N = (3 do 5) x M x n, gdje je M broj razreda,

n dimenzionalnost vektora značajki

Primjer: Sustav za raspoznavanje osoba na temelju lica

npr. 550 osoba, dimenzija vektora značajki n = 110

N = (3 do 5) x M x n = 3 x 550 x 110 = 181 500 uzoraka

Skup uzoraka za učenje i skup uzoraka za ispitivanje

Primjer: Klasifikator brojčano-slovčanih znakova

M = 10 + 30

n = 18

N = (3 do 5) x M x n = 5 x 40 x 18 = 3600 uzoraka