Veštačka inteligencija 1

Kolokvijum 2

26. januar 2024.

Zadatak

Dat je skup podataka za klasifikaciju bolesti srca. Skup se sastoji od kolona kao što su: pol, godine, holesterol, maksimalni otkucaji srca, itd. Labela je da li osoba ima bolest srca ili ne. Potrebno je izanalizirati i pripremiti podatke, a potom i istrenirati traženi model za klasifikaciju. U nastavku su dati opisi kolona u skupu podataka (iz originalnog izvora):

Attribute Information

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS ¿ 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of ¿ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

Odgovore na pitanja možete ostaviti u komentarima u kodu ili u markdown ćelijama.

Napomena: Pri podeli skupa podataka i korišćenju svakog modela, ili bilo koje druge operacije koja ima randomizaciju, potrebno je postaviti seed na 42.

1. Analiza skupa podataka **[7 bodova]**
   - Učitati i prikazati skup podataka.
   - Utvrditi broj instanci i broj kolona u skupu podataka

- Koji atributi su numerički, a koji kategorički? Posebno, za kategoričke atribute utvrditi koji su binarni, a koji višeklasni.
- Iscrtati grafik distribucije godina (kolona Age). Koja je maksimalna vrednost, a koja minimalna za broj godina?
- Izdeliti godine na grupe (bins) po 10 godina (0-10, 10-20, ...) - za one dekade za koje postoje podaci. Potom za svaku grupu izračunati prosečan broj pacijenata koji su oboleli od bolesti srca. Rezultate prikazati u obliku bar grafikona.

2. Priprema skupa podataka [**5 bodova**]
   - Pretvoriti binarne kategoričke atribute u numeričke (0 i 1)
   - Pretvoriti višeklasne kategoričke atribute u numeričke (one-hot encoding)
   - Podeliti skup podataka na atribute i labelu
   - Podeliti skup na trening i test sa odnosom 67:33%
   - Normalizovati atribute (oduzeti minimalnu vrednost i podeliti sa razlikom maksimuma i minimuma) za oba skupa podataka

3. Treniranje osnovnog (baseline) modela [**5 bodova**]
   Potrebno je istrenirati logističku regresiju na trening skupu sa podrazumevanim skupom parametara u scikit'learn biblioteci. Izračunati njegovu tačnost na trening i test skupu. Da li je model overfit-ovan ili underfit-ovan ili ništa od ta dva?

4. Treniranje random forest modela [**5 bodova**]
   Potrebno je istrenirati random forest model na trening skupu sa pri čemo je ćemo menjati hiperpodatak koji se odnosi na minimalan broj instanci koje su potrebne da bi se jedan čvor u stabl podelio. Za moguće vrednosti uzeti 2, 8, 16 i 64. Uporediti tačnosti na trening i test skupu za sve četiri vrednosti. Šta možemo da zaključimo da se dešava sa modelom kada povećavamo ovaj hiperparametar?

5. Podešavanje hiperparametara za random forest model [**3 boda**]
   U ovom delu zadatka sami izaberite hiperparametre koje ćete menjati za random forest model. Cilj je da se dobije model koji ima tačnost na test skupu veću od 90%. Zabeležiti za koje vrednosti hiperparametara se dobije najbolji rezultat.

Napomena: Moguće je da zbog randomizacije u algoritmima sa nekim od gore navedenih modela već prebacite 90% tačnosti na test skupu. U tom slučaju potrudite se da napravite model koji je bolji od tog (nije važno koliko).