Veštačka inteligencija 1

Domaći 2

20. januar 2024.

Zadatak

Dat je skup podataka za klasifikaciju vrste semena bundeve, na osnovu raznih izmerenih parametara semena. Postoje dve klase semena bundeve koje je potrebno predvideti: Cercevelik i Urgup Sirvisi. Odgovore na pitanja možete ostaviti u komentarima u kodu ili u markdown ćelijama.

1. Priprema i analiza skupa podataka **[1 bod]**
   - Učitati i prikazati trening i test skup podataka.
   - Izlistati sve kolone u skupu podataka i utvrditi broj instanci
   - Proveriti da li postoje null vrednosti
   - Da li postoje kategorički atributi u skupu podataka?
   - Koliko je zastupljena svaka od klasa u skupu podataka? Da li je skup podataka balansiran? Da li je bezbedno koristiti tačnost (accuracy) kao metriku za evaluaciju modela?
   - Konvertovati labelu u numerički oblik (0 i 1)
   - Iscrtati grafik distribucije površine (area) i obima (perimeter)
   - Podeliti oba skupa podataka na atribute i labelu
   - Normalizovati atribute (oduzeti srednju vrednost i podeliti sa standardnom devijacijom) za oba skupa podataka
   - Podeliti trening skup na trening i validacioni sa odnosom 75:25%
   - Istrenirati Naive Bayes model (sa podrazumevanim parametrima) na trening skupu i ispitati tačnost (accuracy) na validacionom skupu. Ovo će nam biti baseline za dalje modele. Da li Naive Bayes overfittuje?
2. Klasifikacija random forest modelom **[1 bod]**
   Potrebno je istrenirati random forest model na trening skupu podataka i ispitati za svaku vrednost hiperparametara precisznost na validacionom

skupu. Uporediti tačnost na trening i validacionom skupu i zaključiti da li model underfituje, overfituje ili ništa od ta dva za sve vrednosti hiperpametara. Od hiperparametara ispitatati sledeće vrednosti:

- maksimalna dubina stabala: 3, 5, 7, 9

Kada se dobije najbolji model po tačnost na validacionom skupu, istrenirati ponovo model nad trening i validacionim skupom i ispitati tačnost na test skupu.

3. Klasifikacija neuralnom mrežom **[1 bod]**
   Kreirati neuralnu mrežu, koja se sastoji od 3 sloja: ulaznog, skrivenog i izlaznog. Ulazni sloj treba da ima 16 neurona, skriveni sloj 16 neurona, a izlazni 1 neuron. Aktivaciona funkcija za ulazni i skriveni sloj je relu, a za izlazni sigmoida. Za optimizaciju koristiti Adam algoritam, a za funkciju gubitka binary crossentropy. Za veličinu batch-a uzeti 16, a za broj epoha 500. Ispitati tačnost na validacionom skupu. U kom trenutku (epohi okvirno) model počinje da overfituje? Da biste to utvrdili iscrtajte kako se menjala greška i tačnost tokom treninga na trening i validacionom skupu. Kolika je tačnost na test skupu?

4. Kreirajte model po izboru **[1 bod]**
   Kreirajte model po izboru, sa hiperparametrima koje podesite sami. Ako je potrebno mogu se i podaci obraditi pre treniranja modela (selekcija atributa, ...), smanjiti overfit (biranje hiperparametara, L2 regularizacija za neuralne mreže) ili se može nekoliko modela iskombninovati u jedan (npr. ensembling - videti link). Cilj je dobiti model koji na test skupu postiže više od 90% tačnosti.
   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).