Treći domaći zadatak iz Tehnika i metoda analize podataka

Danilo Veljović, broj indeksa 1120

Opis skupa podataka

Skup podataka koji će se koristiti je skup podataka o biljci perunici. Ovaj skup podataka od atributa sadrži dužinu i širinu čašičnog listića, kao i dužinu i širinu latice ove biljke. Uz ova četiri atributa takode sadrži i informacije o vrsti perunike, odnosno labelu klase kojoj instanca pripada. Tri moguće klase, odnosno vrste perunike, kojima instance mogu da pripadaju su: Iris setosa, Iris virginica i Iris versicolor. Skup podataka sadrži po 50 instanci svake klase, odnosno ukupno 150 instanci. Dataset se može učitati iz biblioteke sklearn, naredbom load_iris().

Način implementacije algoritma

U metodi fit() se vrši treniranje modela. Najpre se izdvoje sve jedinstvene klase u koje mogu instance da se klasifikuju. Nakon toga se inicijalizuju srednje vrednosti, varijanse i apriori verovatnoće za sve klase i atribute na 0. Apriori verovatnoća je verovatnoća koja se računa pre upliva novih podataka. Ona predstavlja frekvenciju pojavljivanja te klase u uzorcima, tj predstavlja odnos između uzoraka koji pripadaju datoj klasi i svih mogućih uzoraka. Zatim se za svaku klasu izdvajaju najpre sve instance trening skupa koje joj pripadaju. Zatim se za svaku od tih klasa računa srednja vrednost, varijansa i apriori verovatnoća.

Metoda kojom se vrši predikcija je funkcija predict() koja interno poziva za svaki element test skupa privatnu funkciju predict. U toj funkciji postoji posteriors[] niz koji čuva posteriori verovatnoće za svaku klasu. Poenta metode predict je da vrati klasu sa najvećom posteriori vrednošću. Ovo se postiže argmax funkcijom koja vraća indeks maksimalnog elementa iz posteriori niza i za taj element se vraća klasa.

Funkcija predict će da iterira kroz sve klase, za svaku od klase da računa posteriori verovatnoću i da je doda u niz posteriori verovatnoća. Posteriori verovatnoća se računa kao zbir logaritama priori verovatnoća i logaritma funkcije gustine raspodele koja se računa u funkciji **pdf()** . Funkcija gustine raspodele je data na slici 1. Funkcija gustine raspodele se računa kao Gausova raspodela.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Slika 1 – Gausova raspodela

Funkcija koja vraća procenjenu verovatnoću je data na slici 2:

$$y = argmax_y \log(P(x_1|y)) + \log(P(x_2|y)) + \ldots + \log(P(x_n|y)) + \log(P(y))$$

Slika 2 – Predikcija klase za instancu test skupa

Performanse modela se vrše cross validacijom. Kros validacija je implementirana ručno u kodu. Performansne mere koje se koriste su preciznost, tačnost, odziv i f1 mera i prikazana je matrica konfuzije.

Kros validacija je testirana na n = 3, tj skup se deli na tri podskupa, dva se koriste za treniranje i jedan za testiranje. Ovaj proces se ponavlja tri puta. Tačnost koja se dobije je: 92.5%, 87.5% i 97.5%. Preciznost nakon kros validacije je 1.0, 88.89%, 88.37%. Odziv je 100%, 87.67%, 89.41%. Model daje dobre performanse za ovu višeklasnu klasifikaciju.