Mašinsko učenje, Jun2 rok, 29. jun 2021.

Na Desktop-u možete pronaći arhivu sa imenom *ML_jun2_2021_materijali.zip* u kojoj se nalaze Jupyter sveske i materijali potrebni za rad. Raspakovati arhivu pa dobijeni direktorijum preimenovati tako da odgovara vašim podacima u formi *ML_jun2_2021_ImePrezime_BrojIndeksa*. Zatim mu pristupiti iz terminala pokretanjem komande *jupyter notebook*.

Na Desktop-u se nalazi i direktorijum sa imenom *docs* u kojem se nalazi dokumentacija.

1. **(10 poena)**

   U datoteci *framingham.csv* se nalaze medicinski podaci stanovnika Framinghama koji su učestvovali u ispitivanjima prevencije kardiovaskularnih bolesti. Skup sadrži 15 atributa koji objedinjuju sociološke i medicinske karakteristike ispitanika kao što su starosna dob, pol, indeks telesne mase, koncentracija šećera u krvi i drugi. Potrebno je napraviti klasifikator koji na osnovu ovih podataka može da predvidi pojavu srčanih problema. Informacije o pojavi srčanih problema ispitanika evidentirane su na nivou atributa *TenYearCHD*.

   a) Učitati podatke koji se nalaze u datoteci *framingham.csv*. Obrisati kolonu *education*, a potom obrisati i sve vrste koje sadrže barem jednu nedostajuću vrednost.

   b) Grafikonom sa stubićima prikazati odnos pacijenata po polu.

   c) Grafikonom sa stubićima prikazati odnos pušača i nepušaća sa srčanim problemima.

   d) Izdvojiti vrednost ciljne promenljive *TenYearCHD* koja predviđa mogućnost oboljenja u narednih 10 godina, a potom podeliti podatke na skup za treniranje i skup za testiranje u razmeri 2:1. Prilikom podele podataka voditi računa o stratifikaciji. Parametar *random_state* postaviti na vrednost 5.

   e) Izvršiti standardizaciju podataka.

   f) Kreirati i obučiti model logističke regresije sa težinama 1 i 4 koje, redom, odgovaraju klasama 0 i 1.

   g) Dati F1 ocenu modela, a potom posebno izračunati i F1 ocene po polovima. Da li model daje bolje predikcije za muškarce ili žene?

   h) Za obučeni model izdvojiti atribute koji imaju najveći pozitivan i najveći negativan uticaj na predikciju.

2. **(12 poena)**

   Potrebno je implementirati autoenkoder koji uči reprezentacije slika skupa *fashion mnist* u kojem se nalaze slike odevnih predmeta.

   a) Koristeći funkciju *load_data* paketa *fashion_mnist* učitati sličice skupa za treniranje i skupa za testiranje.

   b) Iz učitanog skupa za treniranje izdvojiti prvih 30000 sličica za treniranje, a potom narednih 10000 sličica za validaciju.

   c) Učitane sličice transformisati u tip *float32*, a potom izvršiti njihovu normalizaciju.

   d) Napraviti autoenkoder mrežu kod koje se enkoder sastoji od ulaznog sloja koji vrši ispravljanje učitane slike i jednog gustog sloja latentne dimenzije 64 sa ReLu aktivacijom, a dekoder od gustog sloja dimenzije 784 sa sigmoidnom aktivacijom i sloja koji vrši transformaciju slike na polaznu dimenziju. Potom mreži pridružiti Adam optimizator i srednjekvadratnu grešku kao funkciju gubitka.

   e) Trenirati mrežu u 10 epoha koristeći za validaciju pripremljene podatke. Parametar *shuffle* postaviti na vrednost *True*. Veličinu paketića za treniranje postaviti na 32.

   f) Prikazati grafik funkcije gubitak u toku treniranja na skupu za treniranje i skupu za validaciju.

   g) Prikazati za proizvoljnu sliku skupa za testiranje kako radi autoenkoder.

   h) Dati ocenu srednjekvadratne greške autoenkodera na skupu za testiranje.

3. **(8 poena)**

U datoteci *corona_ tweets.csv* se nalaze tvitovi na temu korona virusa prikupljeni sa različitih lokacija. Svakom tvitu je pridruženo obeležje sentimenta, od jako negativnog do jako pozitivnog. Potrebno je generisati 2D prikaz skupa zbog uvida u strukturu skupa i dalje tekstualne obrade.

a) Učitati tvitove koji se nalaze u datoteci *corona_ tweets,* a potom izdvojiti kolone koje se odnose na lokaciju (*Location*), sadržaj tvita (*OriginalTweet*) i sentiment (*Sentiment*).

b) Izdvojiti tvitove koji se odnose na Ameriku i London - njima odgovaraju lokacije *United States* i *London, England*.

c) Prikazati grafikonom sa stubićima učestalost tvitova po sentimentu.

d) Kreirati TfIdVectorizer tako da eliminiše stop reči tipične za engleski jezik i uzima u obzir reči sa barem 4 pojavljivanja. Koliko reči će biti zadržano?

e) Koristeći kreirani vektorizator, pripremiti tviter porukice

f) Koristeći PCA analizu izvršiti redukciju pripremljenih tvitova na dve dimenzije. Koji udeo varijanse je zadržan ovom transformacijom?

g) Pridružiti boje labelama sentimenata prema zadatoj mapi, a zatim prikazati 2D grafik tvitova po sentimentima.

h) Koliki treba da bude broj komponenti PCA analize da bi se zadržalo 90% varijanse polaznog skupa tvitova?