Mašinsko učenje, ispitni rok Jun, 11.06.2022.

Na Desktop-u možete pronaći arhivu sa imenom *ML_jun_2022_materijali.zip* u kojoj se nalaze materijali potrebni za rad. Raspakovati arhivu pa dobijeni direktorijum preimenovati tako da odgovara vašim podacima u formi *ML_jun_2022_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. **(11 poena) Lienarna regresija cena nekretnina u Parizu**

   U datoteci *data/ParisHousing.csv* nalaze se podaci o nekretninama (zgradama) u gradu Parizu. Potrebno je kreirati model linearne regresije koji predviđa cene nekretnina (eng. *price*).

   a) Učitati podatke koji se nalaze u datoteci *data/ParisHousing.csv*, a zatim tabelarno prikazati prvih 15 instanci podataka.

   b) Prikazati statistike (prosečnu vrednost, standardnu devijaciju, minimim, maksimum, 25%, 50% i 75% percentil) numeričkog atributa *squareMeters* i ciljne promenljive *price*.

   c) Grafikonom sa stubićima prikazati odnos broja luksuznih i standardnih nekretnina u skupu podataka.

   d) Analizirati atribut *hasPool* čije su vrednosti obeležene sa 0 i 1 i koje daju informaciju o tome da li uočena zgrada sadrži bazen ili ne. Grafikonom u obliku pitice (eng. *pie chart*) prikazati odnos broja instanci sa i bez bazena. Delove pitice obeležiti procentima.

   e) Iz skupa podataka izbaciti atribute koji se odnose na poštanski kod opštine u kojoj se nekretnina nalazi (eng. *cityCode*) kao i na rang kvarta nekretnine (eng. *cityPartRange*)

   f) Izvršiti pripremu kategoričkog atributa: kategorija nekretnine (eng. *category*: luksuzna ili standardna nekretnina).

   g) Izdvojiti vrednost ciljne promenljive (eng. *price*) koja predstavlja cenu nekretnine (u dolarima), a potom podeliti podatke na skup za treniranje i skup za testiranje u razmeri 2:1. Parametar *random_state* postaviti na vrednost 1989. Prikazati dimenzije kreiranih skupova za trening i testiranje.

   h) Izvršiti standardizaciju podataka svodjenjem proseka na 0 i standardne devijacije na 1.

   i) Kreirati i obučiti model linearne regresije koji predviđa cene nekretnia u dolarima na osnovu 16 zadržanih atributa.

   j) Dati ocenu modela u terminima apsolutne i srednjekvadratne greške, kao i ocenu količine objašnjene varijanse.

   k) Izdvojiti tri najinformativnija atributa kreiranog modela.

2. **(11 poena) Duboka konvolutivna neuronska mreža za klasifikaciju**

   Znakovni jezici jesu jezici koji koriste komunikaciju pokretima za prenošenje značenja. Ovo može uključivati istovremeno korišćenje gestova ruku, pokreta, orijentacije prstiju, tela ili izraza lica kako bi se prenele ideje govornika.

   Skup podataka koji se nalazi u folderu *data/signLanguageDigits/* sadrzi slike u rezoluciji 64x64 na kojima su prikazani jednocifreni brojevi (0-9) (kroz razne poze šake).

   Kreirati klasifikacioni model konvolutivne duboke neuronske mreže sposobne za klasifikaciju brojeva na slikama.

   a) Učitati podatke kao numpy nizove iz foldera *data/signLanguageDigits/*. Koristiti funkciju *load* biblioteke numpy. Prikazati dimenzije ucitanih skupova atributa odnosno labela.

   b) Prikazati prvih 10 instanci ucitanog skupa podataka.

   c) Proširiti skup atributa dodajući mu četvrtu dimenziju koja odgovara monohromatskim slikama (eng. *channel last* notacija). Prikazati dimenziju matrice atributa nakon ove transformacije.

   d) Izvrsiti stratifikovanu podelu skupa podataka u skup za treniranje i skup za testiranje (u odnosu 2:1). Parametar *random_state* postaviti na vrednost 1989. Prikazati dimenzije kreiranih skupova za trening i testiranje.

   e) Prikazati histogram klasa u skupovima za trening i testiranje. Voditi računa o činjenici da su učitane labele već enkodirane.

f) Napraviti model zasnovan na dubokoj konvolutivnoj neuronskoj mreži čiji je opis predstavljen na slici ispod.

Mreza se sastoji od dva konvolutivna sloja sa po 16 filtera praćena (eng. *max pooling*) agregacijom sa veličinom polja 2x2. U konvolutivnim slojevima, koriste se elu aktivacione funkcije, veličina kernela 3x3, i zadržavaju se dimenzije slike (eng. *same padding*).

Poslednji sloj koristi *softmax* aktivacionu funkciju. Vrednost verovatnoća isključivanja neurona (u eng. *dropout* sloju) iznosi 0.15. Broj neurona u dva gusta sloja jeste 64 odnosno 10 redom.

Nakon kreiranja modela, pridružiti mu kategoričku unakrsnu entropiju (eng. *categorical crossentropy*) kao funkciju gubitka zatim ga kompajlirati korišćenjem *Nadam* optimizatora sa korakom učenja 0.001.

g) Obučiti mrežu koristeći paketiće veličine 512 u ukupno 25 epoha. Uspešnost treniranja pratiti koristeći validacioni skup koji predstavlja 20% skupa za trening.

h) Prikazati grafik promena funkcije gubitka i preciznosti u toku treniranja.

i) Izračunati tačnost modela na skupu za testiranje. Prikazati matricu konfuzije na test skupu.

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 64, 64, 16)        160

max_pooling2d (MaxPooling2D  (None, 32, 32, 16)        0
)

conv2d_1 (Conv2D)            (None, 32, 32, 16)        2320

max_pooling2d_1 (MaxPooling  (None, 16, 16, 16)        0
2D)

flatten (Flatten)            (None, 4096)              0

dense (Dense)                (None, 64)                262208

dropout (Dropout)            (None, 64)                0

dense_1 (Dense)              (None, 10)                650

=================================================================
Total params: 265,338
Trainable params: 265,338
Non-trainable params: 0
_________________________________________________________________
```

3. **(8 poena) PCA na skupu podataka Iris**

Implementirati analizu glavnih komponenti (eng. PCA) u cilju redukcije dimenzionalnosti *Iris* skupa podataka.

a) Učitati skup podataka *data/Iris.csv*. Koristiti metod *read_ csv* biblioteke *pandas*. Kao koplonu za indeksiranje koristiti kolonu *Id*.

b) Prikazati statistike učitanih atributa: prosek, standardnu devijaciju, minimum, maksimum, zatim 25, 50 i 75 percentil.

c) Izvršiti mapiranje ciljne promenljive (eng. *Species*) u numeričke vrednosti.

d) Prikazati histograme svakog od četiri atributa kao i raspodelu ciljne promenljive (eng. *Species*)

e) Podeliti podatke na matricu atributa (X) i vektor ciljne promenljive (y).

f) Izvršiti skaliranje atributa korišćenjem klase *MinMaxScaler* biblioteke *sklearn*.

g) Korišćenjem analize glavnih komponenti (eng. *PCA*) modifikovati polazni skup atributa preslikavanjem u prostor minimalne dimenzionalnosti koji čuva bar 90% varijanse podataka.

Prikazati dimenziju novog skupa atributra.

h) Grafikonom sa stubićima (eng. *bar plot*) prikazati količinu varijanse podataka u svakoj od glavnih komponenti *PCA* algoritma.