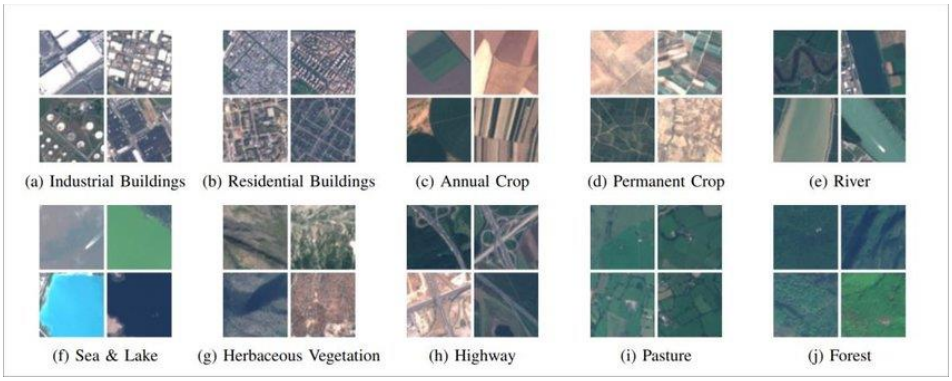
Uvod u znanost o podacima – ispitni zadaci za vježbu

Pitanja vezana uz skupove podataka (praksa i teorija):

1. Skup podataka „Estonia“ sadrži zemlju podrijetla, ime, spol (M muški, F ženski), godine, kategoriju (P putnik, C posada) i sudbinu (0 poginuo, 1 preživio) 989 putnika na broju MS Estonia u noći potonuća. Skup je nešto reduciran i izmijenjen za potrebe ovog zadatka (ovo ne uzeti u obzir za d. podpitanje). Za ovaj skup podataka odredite (estonia-passenger-list_mod_1.csv):
    a. Kako biste klasificirali zadani izvor podataka prema strukturi?
    b. Objasnite o kojoj se vrsti istraživanja prema vrsti korištenih podataka ovdje radi.
    c. Objasnite radi li se ovdje o studiji presjeka ili longitudinalnom istraživanju.
    d. Kako je u ovo istraživanje uključeno 989 putnika, radi li se o slučajnom, pristranom ili neslučajnom uzorku? Objasnite svoj odgovor. Definirajte svaki od navedenih vrsta uzoraka.
    e. Programski izbacite stupce “PassenderId”, “Firstname” i “Lastname”. Za svaki od stupaca “Country”, “Sex” i “Category” iskoristite prikladni enkoder kako biste podatke transformirali u numeričke vrijednosti.
    f. Provedite zamjenu nedostajućih vrijednosti značajke “Age” mjerom očuvanja sredine. Pojasnite koju mjeru sredine koristite i zašto.
    g. Provjerite prikladnim prikazom je li distribucija značajke “Age” simetrična. Ako nije, objasnite o kojoj se vrsti asimetričnosti radi.
    h. Provedite postupak DBSCAN kako biste otkrili stršeće vrijednosti u ovom skupu podataka ne uzimajući u obzir ciljnu značajku (“Survived”). Po potrebi varirajte parametre algoritma te na temelju odgovora algoritma provjerite naznačene pojedince. Navedite redne brojeve stršećih.
    i. Provjerite mogu li se stršeće vrijednosti otkriti koristeći samo metode vizualizacije podataka. Pritom primijenite prikladni graf za njihovo otkrivanje. Koji graf koji ste isprobali daje najbolji prikaz?
    j. Korištenjem prikladnog algoritma nadziranog strojnog učenja s jasnim tumačenjem ustanovite koja zemlja je bila najsretnija pri preživljavanju. Navedite o kojoj se zemlji radi i gdje u modelu se ta činjenica vidi. Pritom model učite na slučajno izdvojenih 67% podataka, a testirajte na preostalom skupu.

2. Skup podataka Possum sadrži podatke o 104 planinska četkasta oposuma pronađena na sedam lokacija u Australiji: oznaka lokacije gdje je pronađen oposum, populacija (Vic – Victoria, other – Novi Južni Wales ili Queensland), spol (M muški, F ženski), dob, duljina glave u mm, širina lubanje u mm, ukupna duljina u cm, duljina repa u cm, duljina stopala u mm, duljina ušne školjke u mm, udaljenost od medijalnog kantusa do lateralnog kantusa desnog oka u mm, obujam prsa u cm i opseg trbuha u cm. Za skup podataka Possum (possum.csv i possum_description.txt):
    a. Navedite metode prikupljanja podataka. Koja od navedenih metoda je mogla biti upotrijebljena prilikom stvaranja ovog skupa podataka? Objasnite.
    b. Kako možemo ukloniti statistički redundantne (nebitne) značajke? Primjenom odgovarajuće metode utvrdite postoje li takve numeričke značajke u ovom skupu i uklonite ih.
    c. Za preostale numeričke značajke iz prethodnog dijela zadatka provedite standardizaciju z-vrijednosti (z-skor) te potom primijenite grupiranje postupkom k-srednjih vrijednosti (uz variranje hiperparametra k između 1 i 5). Pritom prikažite grupiranja primjeraka u dvije dimenzije za parove varijabli iz skupa eye, chest i belly. Uočavaju li se grupe? Na što treba obratiti pozornost pri primjeni algoritma k-srednjih vrijednosti?
    d. Izračunajte srednju vrijednost, standardnu devicijaciju i interkvartalno raspršenje za sve numeričke značajke u ovom skupu. Vizualizirajte ih na istom grafu koristeći kutijasti graf (engl. box plot).
    e. Modelirajte varijablu “totlngth” na temelju ostalih numeričkih varijabli u skupu koristeći višestruku linearnu regresiju pri čemu je prethodno potrebno centrirati prediktorske varijable. Prikažite rezultate modela. Koje su pretpostavke regresijskog modela?
    f. Definirajte srednju kvadratnu pogrešku i primijenite ju na model iz prethodnog pitanja koji se ispituje na izdvojenom testnom skupu (30% primjeraka skupa).

3. EuroSAT je skup satelitskih RGB slika dimenzija 64x64 koje su raspoređene u deset klasa ovisno o tome što je na njima prikazano (urbana područja, šume, polja, …). Vaš zadatak je složiti model dubokog učenja i naučiti ga na danom skupu podataka. Prikažite implementirani kod. (eurosat.zip):



   a. Jedan od prvih koraka u pripremi slika za duboko učenje, je njihova anotacija. Objasnite što je anotirano u danom skupu slika.
   b. Učitajte dani skup podataka i podijelite ga na skup za učenje i skup za testiranje. Za samo učitavanje podataka možete iskoristiti već implementirani razred *torch.utils.data.Dataset*.
   c. Složite model (bilo kakav) kojim ćete izvršiti klasifikaciju danih slika.
   d. Prikažite matricu zabune. Izračunajte točnost, preciznost, odziv (recall) i F1-mjeru (F1-score). Objasnite dobivene rezultate i uspješnost modela kojeg ste upotrijebili.
   e. Napišite funkciju za učenje modela i naučite implementirani model (dovoljno je da model naučite na samo jednoj epohi => *num_epochs=1*). Ispišite iznos funkcije gubitka i točnost vašeg modela na testnom skupu podataka.

Dodatna teorijska pitanja:

4. Radili ste na nekom NLP (Natural Language Processing) problemu i pri učenju modela učili ste i vektorske reprezentacije riječi (word embedding). Nakon svega, odlučili ste baciti pogled na to kako izgledaju te vektorske reprezentacije. Konkretno promatrate reprezentacije sljedećih riječi:
    a. Piramida
    b. Jezikoslovlje
    c. Čavao
    d. Konj
    e. Računalo
    f. Lingvistika
    g. Planet
    h. Knjiga
    i. Kvaka

    Za vektorske reprezentacije kojih dviju riječi (od prethodno navedenih) bi bilo za očekivati da su relativno blizu (euklidska udaljenost)? Zašto? Kolika bi ta udaljenost bila da smo koristili "one-hot encoded" reprezentacije?

5. Marko je pri učenju modela primjetio da mu se model dosta lagano prenauči i stoga je odlučio uvesti određene metode regularizacije. Točnije, odlučio je koristiti "dropout" i to na takav način da bi poredao više "dropout" slojeva jedan za drugim:

```
def Model(nn.Module):
def __init__(self, . . .):
. . .
self.dropout1 = nn.Dropout(p=0.1)
self.dropout2 = nn.Dropout(p=0.15)
self.dropout3 = nn.Dropout(p=0.2)
. . .
def forward(self, x):
. . .
x = self.dropout1(x)
x = self.dropout2(x)
```

```
x = self.dropout3(x)
. . .
```

Ima li ovakvo redanje “dropout” slojeva smisla? Je li moguće postići isti regularizacijski učinak s manje linija koda? Obrazložite svoj odgovor i modificirajte Markov kod!

6. Osim primjene metode očuvanja mjere sredine, koje ste još postupke mogli primijeniti za zamjenu nedostajuće vrijednosti nekom drugom u 1. zadatku?

7. Objasnite princip učenja značajki (engl. feature learning, representation learning) u kontekstu ekspertnih značajki i ulaznih podataka.

8. Objasnite koja je razlika između reziduala i pogreške kod regresije.

9. Što testiramo hi-kvadrat testom, kakve frekvencije koristimo u izračunu hi-kvadrat vrijednosti?

10. Za algoritam k-NN, navedite kako izabrati k u praksi.

11. Navedite koja svojstva funkcija mora zadovoljavati da bude mjera udaljenosti.