Neurális hálózat házak árának predikciójára

3. Házi feladat

2023. november 20.

1. Feladat

Ebben a házi feladatban a cél házak árának predikciója lesz neurális hálózat segítségével. Az eredeti adathalmaz innen származik. Itt található róla egy rövid leírás. A neurális hálózat segítségével a környékre jellemző különböző 8 jellemző alapján (medián jövedelem, házak korának átlaga, szobák átlagos száma, stb.) a házak értékének mediánját kell megbecsülni. A cél, hogy minél pontosabb legyen a becslés, amit MSE (mean squared error) alapján határozunk meg.
Ehhez a megfelelő adathalmazt a moodle-ről lehet letölteni. Az adathalmaz linkje a moodle-ben a házi feladat alatt található "Dataset assignment" alatt található a "Feedback"-nél.
`https://edu.vik.bme.hu/mod/assign/view.php?id=115626`
Ugyanitt olvasható egy "mse range" intervallum is, amely a maximális és minimális pontszámhoz tartozó MSE értékeket jelöli. A kapott pontszámot ezek között lineáris, ezen kívül pedig nearest neighbor interpolációval határozzuk meg.
A feladat tetszőleges környezetben megoldható. Mivel a területen a python a legelterjedtebb, ezért ehhez adunk segítséget.

2. Beadandó

A moodle felületre egy .csv fájlt kell feltölteni a prediktált értékekkel, tehát soronként egy-egy float értéket fog tartalmazni. Ezt python környezetben a numpy segítségével a `np.savetxt` függvénnyel tehetjük meg könnyen.

```
np.savetxt('housing_y_test.csv', y_test, delimiter=",", fmt="%g")
```

Tehát soronként egyetlen kimenetet kell tartalmaznia az eredmény .csv fájlnak:

```
2.611
0.22
3.4701
0.35
3.61
0.267
0.6
1.245
4.20
```

3. Hasznos tudnivalók

3.1. Megoldási környezet

A javasolt megoldási környezet a google colaboratory, ahol iPython notebookokat lehet futtatni. Ennek a nyitólapján található egy leírás a környezetről. Előkészítettünk továbbá egy ilyen notebookot, amely végigvezet a neurális hálózatok elméleti és gyakorlati részletein.

`https://colab.research.google.com/drive/1F7fGUrhx_-PWU9i0dKRzzbYYPL4tAe-1?usp=sharing`
Erről a notebookról először a saját google drive-ba kell egy másolatot készíteni, és csak ezután érdemes futtatni (különben a módosítások elvesznek). Javasolt továbbá az alábbi python könyvtárak használata:

- numpy: vektor- és mátrixműveletek
- tensorflow.keras: neurális hálózatok létrehozása és tanítása
- matplotlib: ábrák készítése
- pandas: tanító- és teszt adatok beolvasása

3.2. Fájlok

A "Dataset assignment" alatt található url-ről letöltött .zip fájlban 3 .csv fájl található:

- housing_x_train_xxxxxx.csv: tanító adathalmaz bemeneti jellemzői
- housing_y_train_xxxxxx.csv: tanító adathalmaz elvárt kimenete
- housing_x_test_xxxxxx.csv: teszt adathalmaz bemeneti jellemzői

Ezek egy pandas (`import pandas as pd`) paranccsal könnyen beolvashatóak:

```
pd.read_csv(f'{file_name}.csv', sep=',', encoding='utf-8').values
```

Ez egy numpy tömböt fog visszaadni, amivel utána könnyű dolgozni.

3.3. Adat előfeldolgozása

A neurális hálózatok könnyebben tanulnak, ha az adat normalizálva van. Ehhez érdemes az `sklearn MinMaxScaler` osztályát használni.
Amit nem szabad elfelejteni, hogy ha a tanító adatot normalizáltuk, akkor a teszt adatot is ugyanazokkal a paraméterekkel kell normlizálni, és a prediktált értékeket vissza kell alakítani fájlba írás előtt.

4. Extra rejtvény - nem a házi feladat része

Helyenként a beadandó.