Strojové učenie a neurónové siete

Zadanie číslo 2

Andrej Dubeň

74657

Obsah

Zadanie

Zadanie č. 2

Pripravené dáta z úlohy č.1 môžu obsahovať veľa rovnakých vzoriek. Cieľom tohoto zadania bude vyhľadať rovnaké alebo podobné vzorky, ktoré sa nachádzajú medzi množinami. Tieto vzorky následne odstránime a tak sa vyheme ich duplicite (Odstránte ich z trénovacej množiny).

Úloha č.1.

Zistite prienik medzi trénovacími, validačnými a testovacími dátami. Myslite na to, že obrazy nemusia byť úplne rovnaké ale aj veľmi podobné. (P.S.: Máte veľké množstvo dát. Dajte si pozor aj na rýchlosť spracovania. Pokiaľ nenájdete úplné riešenie pracujte s menšou mnoužinou dát.)

Úloha č.2.

Nájdite podobné skupiny znakov pomocou zhlukovacích algoritmov K-means, DBSCAN a SOM. Po nájdení skupiny podobných znakov (zhlukov), vykreslite priemerný obraz a obraz najbližší priemernému obrazu. Pre každý nájdený zhluk vypíšte počet prvkov v zhluku a štatistickú odchýlku celého zhluku. DBSCAN musí vyhľadať aspoň 10 zhlukov pričom každý zhluk musí obsahovať apsoň 30 prvkov. Tieto algoritmy spustite aspoň na vašich validačných a testovacích dátach, ktoré budú mať aspoň 10 000 prvkov.

Riešenie

Úloha 1.

Vymazal som rovnaké a podobné obrázky medzi datasetmi z trénovacieho datasetu. Postup je nasledovný:

- Vytvoril som set() lebo nebude obsahovať duplicitné haše
- Zahašoval validačné dáta a pridal ich do setu
- Zahašoval testovacie dáta a pridal ich do setu
- V cykle hašhoval trénovacie dáta a hneď aj porovnával so setom, ak sú zhodné tak som si do arrayu pridal ich index
- Vymazal som dáta podla indexov zhodných dát

Z 530k obrázkov môj algoritmus vymazal 357

```
Total images in training dataset:  370399   370399
Total images in testing dataset  79360
Total images in validation dataset  79360
Total number of images in final pickle file   529119
C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_large/Final.pickle  already exists
Removed images from training data: 14992
```

Úloha 2.

V tejto úlohe som naimplementoval zhlukovacie algoritmi Kmeasns, DBScan a SOM. Každý algoritmus vypíše počet dát v zhluku a ich štaticstickú odchíľku, vykreslí priemerný obraz a najbližší obraz k priemernému obrazu.

Kmeans

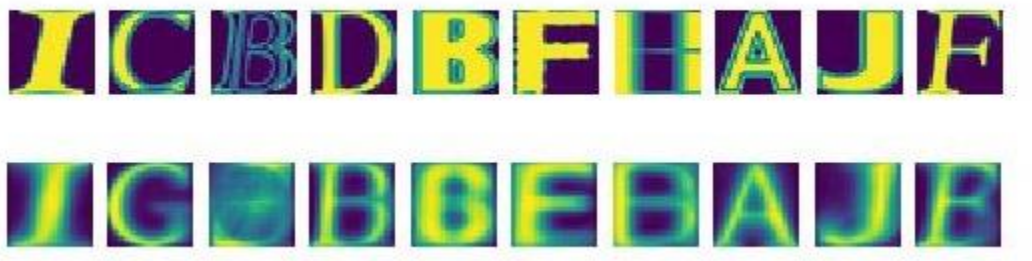
Na vytvorenie zhlukov som použil knižnicu: sklear.cluter.Kmeans a jej metódy:

- KMeans(n_clusters) – počet zhlukov nastavený na 10
- Kmeans.fit(dataset) – dataset bolo potrebné zmeniť na 2D (index, plocha)

Výsledné zhluky:

```
Cluster0 have length: 689 and std: 0.465856
Cluster1 have length: 1259 and std: 0.437697
Cluster2 have length: 1968 and std: 0.347885
Cluster3 have length: 1009 and std: 0.452599
Cluster4 have length: 1235 and std: 0.387796
Cluster5 have length: 646 and std: 0.46601
Cluster6 have length: 766 and std: 0.46503
Cluster7 have length: 686 and std: 0.448646
Cluster8 have length: 639 and std: 0.461341
Cluster9 have length: 1103 and std: 0.442002
```

Zobrazenie priemerných obrazov(dole) a najbližších k nim (hore):



DBScan

Na vytvorenie zhlukov som použil knižnicu: sklear.cluter.DBScan a jej metódy:

- DBScan(eps, min_samples) – eps je vzdialenosť medzi dátami, min_samples – minimálny počet dát v zhluku
- DBSCAn.fit(dataset) – dataset bolo potrebné zmeniť na 2D (index, plocha)

Výsledné zhluky:

```
Cluster0 have length: 68 and std: 0.444706
Cluster1 have length: 38 and std: 0.133298
Cluster2 have length: 51 and std: 0.406531
Cluster3 have length: 53 and std: 0.451202
Cluster4 have length: 62 and std: 0.474434
Cluster5 have length: 45 and std: 0.436144
Cluster6 have length: 53 and std: 0.456101
Cluster7 have length: 31 and std: 0.420686
Cluster8 have length: 39 and std: 0.440914
Cluster9 have length: 30 and std: 0.457958
```

Zobrazenie priemerných obrazov(dole) a najbližších k nim(hore):

SOM

Na vytvorenie zhlukov som použil knižnicu: MiniSom a jej metódy:

- MiniSom(počet neurónov, počet neurónových vrstiev, plocha, počiatočné rozloženie, počiatočný training_rate)
- Train_random(dataset, iterations) – dáta náhodne preusporiadané
- Train_batch(dataset, iterations) – dáta usporiadané podla usporiadania datasetu
- Winner(data) – vráti výherný neurón pre daný obrázok

Výsledné zhluky:

```
Cluster0 have length: 3491 and std: 0.400935
Cluster1 have length: 1856 and std: 0.419139
Cluster2 have length: 1611 and std: 0.45364
Cluster3 have length: 879 and std: 0.324042
Cluster4 have length: 476 and std: 0.442761
Cluster5 have length: 892 and std: 0.462054
Cluster6 have length: 1137 and std: 0.468865
Cluster7 have length: 683 and std: 0.455099
Cluster8 have length: 590 and std: 0.441911
Cluster9 have length: 1511 and std: 0.450102
```

Zobrazenie priemerných obrazov(dole) a najbližších k nim(hore):