Zadanie č. 2

Algoritmy strojového učenia obsahujú veľa hyperparametrov, ktoré sa nastavujú bez exaktných pravidiel a s ohľadom na vstupné dáta. Predtým, než začnete aplikovať prvé algoritmy strojového videnia je preto potrebné poznať svoje dáta. V rámci ďalších zadaní sa budete snažiť, aby váš systém určoval fotografie rovnakých ovocí ako podobné a iných ovocí ako odlišné - preto je dobré vedieť, nakoľko sú si dáta podobné pixel po pixeli v rámci jednotlivých tried aj medzitriedne.

Úloha č.1.

Cieľ: zistite, nakoľko sú si podobné dáta v rámci jednotlivých tried

Vstup: uložené binárne dáta pre jednotlivé ovocia

Postup: načítajte dáta z predošlého zadania (uložené podľa jednotlivých tried). Porovnajte jednotlivé vzorky medzi sebou:

A) vzdialenosťou pixelov (L2, Manhattan)

C) vzhľadom na ich vzdialenosť od priemernej vzorky pre danú triedu

Analyzujte získané výsledky - ktorá trieda je najviac rozmanitá, ktorá najmenej (odporúčame štatistickou analýzou výsledkov). Prenesú sa tieto vlastnosti aj po vašom delení na trénovacie/validačné/testovacie množiny?

Výstup: zrozumiteľné zobrazenie výsledkov (odporúčame formou grafov a základných štatistík). Postup aj získané grafy dobre opíšte v dokumentácii.

Úloha č.2.

Cieľ: zistite, nakoľko sú si podobné dáta v rozdielnych triedach

Vstup: uložené binárne dáta pre trénovaciu množinu (testovaciu, validačnú).

Postup: načitajte dáta z predošlého zadania (uložené do množín pre ďalšiu prácu). Zistite, ktoré vzorky (triedy) sú si v rámci tejto množiny navzájom najpodobnejšie. Využívať na to budete zhlukovacie algoritmy:

A) k-means clustering

B) DBSCAN/ Chinese whispers

C) SOM (za dva bonusové body),

ktoré vám vrátia skupiny(zhluky, clustre) navzájom podobných obrázkov v rámci vášho datasetu. Pre každý získaný zhluk nakoniec zobrazte centrum zhluku (priemerný obrázok) a vypíšte, aké je triedne zloženie pre daný zhluk (v prípade veľkého množstva

zhlukov zobrazte len vhodnú podmnožinu). Pokúste sa analýzou týchto dát zodpovedať na otázku, ktoré ovocia sú si navzájom najpodobnejšie.

Výstup: zrozumiteľné zobrazenie výsledkov - príklady zobrazenia centier aj príslušníkov niekoľkých zhlukov a ku nim prislúchajúce triedne zloženie. Postup aj obrázky dobre opíšte v dokumentácii. Uveďte, ktoré ovocia sú si najviac podobné a ako ste dospeli ku takému záveru.

Poznámky:

- Opäť sa môžete stretnúť s problémami s výpočtovou náročnosťou alebo s nedostatočnou pamäťou. Odporúčame preto najprv si skúšať vaše algoritmy na malej podmnožine dát a až potom na celom datasete. V prípade potreby môžete zmenšiť počet vašich dát (rozumne) alebo použiť knižničnú funkciu na zmenšeniu rozmeru porovnávaných matíc (napr. OpenCV resize).

- Ktoré štatistické ukazovatele sú vhodné, postup analýzy zhlukov a čo je zrozumiteľné zobrazenie nechávame na vás. Malo by však byť z dokumentácie jasné, ku akým záverom ste dospeli a na základe akých podkladov. With great power comes great responsibility.

Zdroje:

- k-means clustering:
  - http://stanford.edu/~cpiech/cs221/handouts/kmeans.html
  - https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
  - https://docs.opencv.org/3.0-beta/doc/py_tutorials/py_ml/py_kmeans/py_kmeans_opencv/py_kmeans_opencv.html
  - https://docs.opencv.org/3.2.0/de/d63/kmeans_8cpp-example.html
  - http://dlib.net/ml.html#kkmeans
- DBSCAN:
  - https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/
  - http://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html
- Chinese whispers:
  - http://dlib.net/ml.html#chinese_whispers
  - https://pdfs.semanticscholar.org/c64b/9ed6a42b93a24316c7d1d6b3fddbd96dbaf5.pdf?_ga=2.69337112.968869931.1539178445-340987529.1539178445