Prípadová štúdia 2

VOS2, 2018/19

Organizácia práce

- Samostatná práca v malých tímoch na zadaných témach
  - veľkosť skupín: 3-4 ľudia, čiže dokopy 4 skupiny (3 skupiny po 4 ľuďoch, 1 s tromi)
- Na každom seminári budú skupiny priebežne reportovať, ako postupujú
- Príklady tém prezentácií (nie nevyhnutne v tomto poradí)
  - Analýza problému (na základe preštudovaných článkov a iných dostupných materiálov) - *mali by ste naštudovanú oblasť vedieť vysvetliť ostatným skupinám*
  - Analýza zvoleného datasetu/datasetov
  - Opis implementácie algoritmov/použitých knižníc
  - Opis vykonaných experimentov/overenia

Hodnotenie

- Report o progrese (v podobe prezentácie alebo notebooku)
  - Každý týždeň (4x10b)
- Ukončenie: odovzdaný report (Jupyter notebook) za 10b
  - v reporte napíšte, kto čo robil
- Hodnotenie spolu: 4x10b za každý kontrolný bod + 10b na záver za odovzdaný report = spolu **50b**

#1 Stávkovanie

- Cieľom je vyhodnotenie úspešnosti rôznych stratégií stávkovania pomocou simulácie
  - Napr.: Triviálne pravidlové stratégie alebo predikcia výsledku zápasu a využitie istoty predikcie ako kurzu
- Dáta - https://www.kaggle.com/hugomathien/soccer/version/10
  - Základná tabuľka je Match. Výsledok a kurzy v rôznych stávkových kanceláriách. Veľa asociovaných tabuliek.
- Zdroje, z ktorých sa dá odraziť
  - Populárne video s príkladom trénovania modelu a odkazmi na ďalšie zdroje. Pozor, robí tam vážnu chybu, a to, že používa štatistiky známe na konci hry na predikciu výsledku hry. Nie veci, ktoré pozná pred samotnou hrou. https://www.youtube.com/watch?v=6tQhoUuQrOw
  - Diplomovka z minulého roku: http://www.itspy.cz/wp-content/uploads/2017/11/IT_SPY_2017_Diplomov_prce_69.pdf

#2 Kryptomeny

- Oplatí sa kupovať kryptomeny, keď ešte len začínajú a majú nízku cenu, veľký potenciál rastu, ale tiež veľkú pravdepodobnosť zániku? Je takáto stratégia rentabilná? Overte túto alebo ďalšie stratégie na historických dátach o vývoji ceny kryptomien.
- Datasetov je veľmi veľa. Väčšinou sú len čiastočné alebo obsahujú len nejakú podmnožinu mien. Tu je jeden veľmi dobrý, ale treba sa pozrieť aj po ďalších: https://www.kaggle.com/jessevent/all-crypto-currencies
- Treba dať pozor na
  - Globálny trend (keď rastia jedna kryptomena, môže to zvýšiť dôveru aj v ostatné).
  - Survivorship bias
- Téma nie je primárne zameraná na krátkodobé špekulatívne obchodovanie (ktorú kedy predať/kúpiť), ale ani toto nie je zakázané.

#3 Zhlukovanie

- Cieľom je naštudovať si rôzne algoritmy zhlukovania, metódy vyhodnocovania kvality zhlukov a ich vizualizácie (PCA, resp. redukcia dimenzionality, dendrogramy a pod.)
- Neobmedzovať sa len na klasické algoritmy (hierarchické, k-means)
- Počas riešenia témy by ste mali zodpovedať nasledovné otázky
  - Aký je state-of-the-art v zhlukovaní?
  - Ktorý zhlukovací algoritmus iný ako k-means by som mal štandardne použiť (ak taký je)?
- Algoritmy demonštrujte na zvolených datasetoch:
  - Napr. Shape sets z http://cs.joensuu.fi/sipu/datasets/
  - Alebo datasety z tejto sady: https://github.com/deric/clustering-benchmark

#4 Učenie súborom metód

- Cieľom je navrhnúť a overiť viacvrstvový ensemble model (modely) využívajúci stratégie učenia súborom metód (resp. ich kombinácie)
- Bude potrebné si pritom naštudovať a vedieť vysvetliť rôzne stratégie učenia súborom metód
  - Bagging, boosting, stacking: https://www.quora.com/What-are-the-differences-between-the-three-commonly-ensemble-learning-techniques-stacking-boosting-and-bagging
  - https://pdfs.semanticscholar.org/5ef3/312c867bb6883cd8c732bdfe89c77e3b2113.pdf
  - https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249
- Algoritmy otestujte na zvolenom datasete
  - Odporúčame vybrať niečo z Kaggle (nejaký Playground/Getting started dataset)

#5 Co-training

- Co-training je príkladom semi-supervised prístupu, ktorý sa používa, ak máme málo označkovaných dát a veľa neoznačkovaných
  - https://dl.acm.org/citation.cfm?id=279962
  - https://dl.acm.org/citation.cfm?doid=354756.354805
  - https://www.sciencedirect.com/science/article/pii/S1566253516302032
- Vašou úlohou je vyskúšať co-training na zvolenom klasifikačnom probléme
- Porovnajte takýto prístup s klasickým supervised prístupom (ktorý sa učí na všetkých črtách, resp. podmnožinách čŕt na označkovaných dátach), resp. základným ensemble prístupom
- Vyhodnoťte úspešnosť modelov, ako aj ich náchylnosť na pretrénovanie
- Odporúčame zreplikovať prácu z prvého článku na inom datasete/datasetoch, napr. https://www.kaggle.com/c/stumbleupon

#6 Optimalizácia hyperparametrov

(Breaking Free of the Grid)

- Cieľom je vyskúšať rôzne stratégie optimalizácie hyperparametrov, opísať, vysvetliť a porovnať na zvolenom klasifikačnom/regresnom probléme (napr. Kaggle)
- https://sigopt.com/blog/breaking-free-of-the-grid/
- Nejaké zaujímavé zdroje:
  - Vysvetlenie Bayesovskej optimalizácie spolu s kódom a príkladmi. https://github.com/fmfn/BayesianOptimization
  - Veľmi pekný blog porovnávajuci rôzne optimalizačné metódy https://towardsdatascience.com/automated-machine-learning-hyperparameter-tuning-in-python-dfda59b72f8a