

Proiecte Data Mining. (2015-2016)

A. Proiecte orientate către algoritmi

Proiectele de tip A constau în:

- Un raport care în care sunt descrise particularitățile problemei abordate, este prezentat cel puțin un algoritm de rezolvare (folosind bibliografia de start și eventual alte lucrări) și sunt prezentate rezultatele obținute aplicând algoritmul implementat (pentru seturi de date simple).
- Implementarea unui algoritm (limbajul de programare este la alegere).

Tematici pentru proiecte de tip A:

1. Algoritmi pentru selecția atributelor (e.g. implementarea algoritmului Relief sau a unui algoritm greedy de tip forward). Biblio: FeatureSelection folder
2. Algoritmi pentru discretizarea atributelor (e.g. implementarea algoritmului Holte 1R). Biblio: FeatureDiscretization folder
3. Algoritmi pentru construirea arborilor de decizie (e.g. implementarea algoritmului ID3 - <http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm>). Biblio: DecisionTree folder
4. Algoritmi de acoperire cu reguli (e.g. implementarea algoritmului PRISM). Biblio: CoveringAlgorithms folder
5. K-Nearest Neighbor (e.g. implementarea algoritmului kNN bazat pe distanța euclidiană). Biblio: kNN folder
6. Clasificator Naïve Bayes (e.g. implementarea unui algoritm pentru date cu attribute discrete). Biblio: NaiveBayes folder
7. Perceptron multinivel antrenat cu Backpropagation (e.g. implementarea unei rețele neuronale feedforward, cu un nivel ascuns și antrenată cu backpropagation – testare pt XOR). Biblio: MLP+BP folder
8. KMeans (e.g. implementarea variantei clasice a algoritmului). Biblio: kMeans folder
9. Fuzzy c-means (e.g. implementarea variantei standard propusă de Bezdek). Biblio: FuzzyCMeans folder
10. Algoritmi aglomerativi de grupare (e.g. implementarea variantei single-linkage variant). Biblio: HierarchicalAlgorithms folder
11. DBSCAN (e.g. implementarea unei variante a DBSCAN). Biblio: DBSCAN folder
12. Algoritmul Apriori (e.g. implementarea unei variante simple a algoritmului Apriori). Biblio: Apriori folder

B. Proiecte orientat inspre date (seturi de date de la UCI Machine Learning Repository)

Proiectele de tip B constau in:

- Un raport in care este descris setul de date, problema care urmeaza a fi rezolvata si metoda utilizata (pe baza lucrarilor mentionate in descrierea setului din UCI Machine Learning Repository)
- Descrierea fluxului de prelucrări (etapele de prelucrare aplicate asupra setului de date), valorile parametrilor si rezultatele obtinute aplicând un instrument de data mining (la alegere – poate fi Weka sau altă platform) to the dataset

Tematici pentru proiecte de tip B:

13. DBWorld e-mails data set (<http://archive.ics.uci.edu/ml/datasets/DBWorld+e-mails>). **Scop:** clasificarea e-mailurilor in 2 categorii: anunturi de conferinte vs alte mesaje (clasificare binară)
14. Microblog PCU data set (<http://archive.ics.uci.edu/ml/datasets/microblogPCU>). **Scop:** indentificarea spammer-ilor (clasificare binară)
15. SMS Spam Collection (<http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection>). **Scop:** clasificarea SMS-urilor in spam/ham (clasificare binară)
16. Energy efficiency data set (<http://archive.ics.uci.edu/ml/datasets/Energy+efficiency>). **Scop:** predicția consumului de energia într-o clădire (pe baza altor caracteristici)
17. GPS trajectories (<http://archive.ics.uci.edu/ml/datasets/GPS+Trajectories>). **Scop:** indentificarea grupurilor de traiectorii similare (clustering)
18. Blog feedback dataset (<http://archive.ics.uci.edu/ml/datasets/BlogFeedback>). **Scop:**predictia numarului de comentarii in urmatoarele 24h (regresie)
19. Online news popularity (<http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>). **Scop:** predicția numărului de partajări ale stiriilor (regresie)
20. Student performance dataset (<http://archive.ics.uci.edu/ml/datasets/Student+Performance>). **Scop:** predicția notei (la matematică, portugheză sau nota finală)
21. AAAI2013 Accepted Papers Dataset (<http://archive.ics.uci.edu/ml/datasets/AAAI+2013+Accepted+Papers>). **Scop:** clustering bazat pe cuvinte cheie
22. News aggregator dataset (<http://archive.ics.uci.edu/ml/datasets/News+Aggregator>). **Scop:** gruparea știrilor pe categorii

C. **Proiecte orientate înspre Scikit-learn (Machine Learning in Python)** - <http://scikit-learn.org/stable/index.html#>

Proiectele de tip C constau în:

- Un raport în care se descrie problema abordată, metodele și modulele din Scikit-learn utilizate pentru a rezolva problema
- Detalii privind utilizarea funcțiilor din Scikit-learn și rezultatele obținute aplicându-le pe setul de date din exemplu și pe un alt set de date

Tematici pentru proiecte de tip C

23. Clasificarea documentelor text utilizând "sparse features" - http://scikit-learn.org/stable/auto_examples/text/document_classification_20newsgroups.html#example-text-document-classification-20newsgroups-py
24. Gruparea documentelor text folosind k-means - http://scikit-learn.org/stable/auto_examples/text/document_clustering.html
25. Completarea imaginii unei fețe - http://scikit-learn.org/stable/auto_examples/plot_multioutput_face_completion.html#example-plot-multioutput-face-completion-py
26. Recunoașterea fețelor folosind "eigenfaces" și SVM - http://scikit-learn.org/stable/auto_examples/applications/face_recognition.html#example-applications-face-recognition-py
27. Recunoașterea cifrelor scrise de mână - http://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html#example-classification-plot-digits-classification-py
28. Discretizarea culorilor folosind K-Means - http://scikit-learn.org/stable/auto_examples/cluster/plot_color_quantization.html#example-cluster-plot-color-quantization-py
29. Discretizare vectorială - http://scikit-learn.org/stable/auto_examples/cluster/plot_face_compress.html#example-cluster-plot-face-compress-py