*Machine Learning* i *Datamining*

Syllabus

Celem zajęć jest zapoznanie słuchaczy z zagadnieniami związanymi z *Machine Learning* i *Datamining*. We współczesnym świecie mamy do czynienia ze zjawiskiem *Data deluge*, czyli zalewu danych. Dane zbierane są w coraz większych ilościach, miliardy urządzeń z każdą chwilą rejestrują wszystko, co tylko się da - temperaturę, napięcie w sieci elektrycznej, natężenie ruchu samochodowego, kursy giełdowe, ilość UV w promieniowaniu słonecznym etc. Zdobycie danych staje się coraz mniejszym problemem, coraz większym problemem jest natomiast sensowne wykorzystanie tych danych.
Dlatego analitycy danych są poszukiwani (i w najbliższej przyszłości będą bardzo poszukiwani) na rynku pracy. Ich rola będzie rosła w miarę, jak procesy decyzyjne w zarządzaniu w coraz większym stopniu oparte będą o inteligentne technologie analizy danych, zwłaszcza tzw. *big data*, czyli danych o dużej objętości. W nowoczesnym zarządzaniu w wielu przypadkach nieodzowne staje się zautomatyzowanie przetwarzania danych oraz procesu decyzyjnego z uwagi na szybkość zachodzących procesów oraz ilość istotnych danych.

Wymagania: Od słuchaczy oczekuje się znajomości matematyki w zakresie algebry liniowej, rachunku prawdopodobieństwa na poziomie podstawowym oraz umiejętności programowania.

1. Co to jest Machine Learning/Datamining?
    - Regresja vs. klasyfikacja
    - Supervised/semi-supervised/unsupervised learning
    - Modele parametryczne i nie-parametryczne
    - Ocena jakości modelu - RMS, ROC, AUC
    - Dane treningowe/testowe, walidacja modelu
    - Cross-validation
2. Omówienie klasycznych algorytmów w zestawie narzędzi górnika danych
    - Regresja liniowa, metoda najmniejszych kwadratów
    - Perceptron
    - Regresja logistyczna, estymator największej wiarygodności
    - Support Vector Machines (SVM)
    - Naive Bayes
    - Drzewa losowe i pochodne
    - Sieci neuronowe
    - KNN
3. Algorytmy selekcji atrybutów
    - Orthogonal Matching Pursuit (OMP)
    - Lasso
    - Selekcja grupowa
    - "Greed is good" --- analiza teoretyczna OMP

4. Klastrowanie
5. Metody polepszania wydajności algorytmów
    - Bagging
    - Boosting
    - Stacking
    - Ensemble
6. Wyciek informacji - leakage
    - Co to jest wyciek informacji? Zawody vs. świat realny.

Wykład zorganizowany w ramach projektu Interdyscyplinarne Studia Doktoranckie ,,Społeczeństwo-Technologie-Środowisko'' współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego.

Harmonogram zajęć (12 godzin wykładu (64 miejsca)) i (6 godzin zajęć na pracowni komputerowej (20 miejsc)), wszystkie zajęcia w budynku Wydziału Matematyki i Informatyki UJ ul. Łojasiewicza 6

poniedziałki: 24 lutego i 3 marca wykład 12-14 sala 0094, pracownia 16-18 sala 0028
środy: 26 lutego i 5 marca, wykład 16-18 sala 0094, pracownia 10-12 sala 0028
piątki: 28 lutego i 7 marca wykład 12-14 sala 0094, pracownia 10-12 sala 0016