QuantUp
Artur Suchwałko
+48-511-175-949
artur@quantup.pl
http://www.quantup.pl

Program szkolenia

Budowa modeli predykcyjnych z wykorzystaniem R

Prowadzący: dr inż. Artur Suchwałko

1. Wprowadzenie
   - zastosowania modeli predykcyjnych
   - przygotowanie danych
   - etapy uczenia i testowania efektywności modelu
   - dobór parametrów modeli
2. Przygotowanie danych
   - analiza pojedynczych cech
     - rozkłady cech (tablice kontyngencji, histogramy)
     - obserwacje brakujące oraz obserwacje odstające
     - kontrola jakości i czyszczenie danych
     - wstępny wybór cech do konstrukcji modelu - analiza zdolności dyskryminacyjnej cech
   - przedziałowanie zmiennych ciągłych (dyskretyzacja)
     - rola przedziałowania
     - metody przedziałowania
       * weight of evidence (WoE)
       * maksymalizacja entropii
       * drzewa klasyfikacyjne
   - analiza zależności między cechami i konstrukcja cech pochodnych (generated characteristics, cross characteristics)
   - standaryzacja
   - próbkowanie
3. Metody klasyfikacyjne i regresyjne
   - analiza dyskryminacyjna
   - metoda najbliższego sąsiada
   - sieci neuronowe
   - maszyny wektorów podpierających (SVM)
   - drzewa klasyfikacyjne
   - drzewa regresyjne
   - randomForest

- klasyfikator Bayesa
- regresja liniowa
- regresja logistyczna

4. Modele oparte na drzewach
   - specyfika modeli opartych na drzewach
   - przegląd zastosowań modeli opartych na drzewach
   - wizualizacja i interpretacja wyników
   - praktyczne aspekty związane z budową modeli opartych na drzewach:
     - kryteria wyboru zmiennych
     - kryteria podziału
     - kryteria zatrzymania
     - ocena złożoności struktury drzewa
   - drzewa klasyfikacyjne
   - drzewa regresyjne
   - postprocessing drzew: upraszczanie i modyfikacje struktury drzew (pruning), analiza ekspercka
   - zalety i wady modeli opartych na drzewach.
   - poprawa stabilności i efektywności drzew (algorytm bagging, modele hybrydowe)
   - lasy losowe (random forest)
5. Ocena jakości klasyfikacji i dobór parametrów klasyfikatorów
   - ocena błędu klasyfikacji
   - ocena jakości modelu: train/test, cross-validation, leave-one-out, bootstrap
   - krzywa ROC, współczynnik AUROC
   - cost-sensitive learning, cost-sensitive evaluation
   - dobór optymalnego punktu odcięcia
   - dobór optymalnych parametrów klasyfikatorów
   - porównanie i wybór najlepszego modelu
6. Wybór cech do modelu
   - kryteria zastosowania cech w modelach (statystyczne, biznesowe, operacyjne)
   - metody graficzne
   - przegląd zupełny zbioru cech
   - metody jednokrokowe (filtry)
   - metody wielokrokowe (forward, backward, forward-backward)
   - metody wbudowane w klasyfikatory (np. randomForest), komitety modeli, inne metody
7. Bardzo ważne praktyczne aspekty modelowania
   - budowa modeli dla małych zbiorów danych
   - budowa modeli dla cech numerycznych (ilościowych) bez przedziałowania
   - zależność cech (numerycznych i kategorycznych) — jak sobie z nią poradzić
   - nierówne proporcje grup i jej konsekwencje
   - porównanie podejść do budowy modeli: dummy variables, przekodowanie WoE, modele dla zmiennych ciągłych
8. Dodatkowe zagadnienia praktyczne związane z budową modeli R
   - formaty danych wejściowych
   - współpraca z MS Excel
   - eksport modeli w formacie PMML