Data mining - ćwiczenia

Marcin K. Dyderski

18 kwietnia 2018

Data mining - zadania

1. Wczytaj plik 'beer.csv' dostępny na githubie. Zawiera on dane o ocenach piw, ich stylu oraz cechach. browar - nazwa browaru nazwa - nazwa piwa BLG - zawartość ekstraktu (%) V - zawartość alkoholu (%) styl - styl piwny IBU - zawartość goryczki centyl - ocena centylowa (10 - oznacza że 10% piw jest gorsze, 90% lepsze) weighted - ocena średnia ważona oceny pochodzą z portalu ratebeer.com
2. Zbiór danych piwnych podziel wg ocen centylowych na zbiór testowy i treningowy używając funkcji caret::createDataPartition()
3. Przygotuj model ctree dla tej cechy, oceń jego dokładność na podstawie RMSE i R2. Przy użyciu funkcji train zawrzyj preprocessing używając argumentu preProcess=c('center',''scale').
4. Przygotuj wizualizację - czy otrzymane wyniki łatwo zinterpretować?
5. Powtórz kroki 3 i 4 dla modeli randomForest (method='rf') oraz gradient boosted modeling (method='gbm'). Który z nich jest najlepszy?
6. wykonaj predykcję modelu dla piwa Halne Mocne: newdata=data.frame(browar='Van Pur'),nazwa='Halne Mocne', BLG=15.1, V=7, style='eurolager', IBU=36).
7. Z githuba pobierz omawiany zbiór danych 'osmunda.csv'. Wybierz jedną z metod i przygotuj model rozmieszczenia gatunku. Zależy nam bardziej na dokładności niż zrozumieniu modelu. Która metoda jest najlepsza?
8. Z githuba ściągnij zbiór danych 'miasto.csv' zawierający cechy funkcjonalne podzbioru 258 gatunków roślin w Poznaniu. Wykonaj model który uważasz za najlepiej nadający się do predykcji stopnia zagrożenia (1. kolumna) na podstawie cech funkcjonalnych. Bardziej zależy nam na zrozumieniu zależności niż na dokładności predykcji.