Eksploracji danych

Lista 5

1. Dla danych readingSkills{party}[R] skonstruuj drzewo decyzyjne na zbiorze treningowym korzystając z kryterium podziału na podstawie zysku informacyjnego i jako atrybut dezycyjny przyjmij mówiący w języku ojczystym (nativeSpeaker) a za atrybuty podziałowe przyjmij wiek (age) i wyniki z testu (score). Podziel losowo zbiór danych na zbiór treningowy (170 rekordów) i zbiór testowy (30 rekordów). Wyznacz błąd treningowy i błąd testowy. Sprawdź czy osoby o cechach (age=11, score=34.45628) i (age=7, score=36.34538) są osobami mówiącymi w języku ojczystym.

2. Wykonaj to samo co w Zad. 1 korzystając z kryterium podziału na podstawie współczynnika podziału Gini'ego. Skorzystaj z funkcji sample.split{caTools} i subset{base} w celu podziału danych na zbiór treningowy i testowy.

3. Załaduj następujące dane mydata=read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv"). Dane zawierają 400 kandydatów do pewnej szkoły z poprzedniego roku i dla danego kandydata pokazują czy kandydat dostał się do danej szkoły i jego wyniki egzaminu do szkoły (gre, punktacja, najwyższa 800), średnia ocen z poprzedniej szkoły (gpa, najwyższa 4.0) i ranking poprzedniej szkoły (rank, 1 nawyższy, 4 najniższy). Zbuduj drzewo klasyfikacyjne korzystając z kryterium podziałowego opartego na zysku informacyjnym dzieląc dane na część uczącą i testową. Wyznacz trafność treningową i testową. Przeprowadź prognozę dla kandydatów z cechami $gre = 700, gpa = 3.7, rank = 2$ i $gre = 650, gpa = 4.0, rank = 1$.

4. Dla danych Carseats{ISLR} skonstruuj i narysuj drzewo regresyjne ze zmienną objaśnianą *Sales* i pozostałymi zmiennymi jako zmienne objaśniające. Przytnij drzewo z optymalnym parametrem złożoności względem błędu *xerror*. Co to jest błąd *xerror*?

5. Dla danych spam7{DAAG} skonstruuj i narysuj drzewo klasyfikacyjne ze zmienną decyzyjną *yesno* i pozostałymi zamiennymi jako zmienne objaśniające. Przytnij drzewo z optymalnym parametrem złożoności

względem błędu *xerror*. Podziel dane na część uczącą i testową. Skonstruuj drzewo klasyfikacyjne na zbiorze uczącym i sprawdź jego trafność na zbiorze testowym podając macierz pomyłek.

6. Podziel dane ames{modeldata} na zbiór treningowy i testowy (70%, 30%). Dla zbioru treningowego skonstruuj i narysuj drzewo regresyjne ze zmienną objaśnianą *Sale_Price* i pozostałymi zmiennymi jako zmienne objaśniające. Przytnij drzewo z optymalnym parametrem złożoności względem błędu *relative error*. Co to jest błąd *relative error*? Podaj prognozę dla zbioru testowego i znajdź błędy RMSE i MASE na zbiorze testowym.