Data mining: projekt indywidualny

Studia Tutorskie

Termin oddania - 10.05.2024

Prezentacja wyników - 14.05.2024

Zadania (15 pkt)

Zad. 1 (3 pkt) Zbiór `Cars93`, znajdujący się w bibliotece `MASS`, zawiera dane dotyczące różnych modeli samochodów osobowych.

(a) Wczytać ten zbiór. Uzyskać bezpośredni dostęp do zmiennych w tym zbiorze. Sprawdzić jakie informacje zawarte są w następujących kolumnach: `Min.Price`, `MPG.city`, `MPG.highway`, `Weight`, `Origin`, `Type`. Utworzyć nowe zmienne opisujące: zużycie paliwa (mierzone w litrach na 100 km) podczas jazdy samochodu w mieście, zużycie paliwa podczas jazdy samochodu na autostradzie, wagę samochodu w kg oraz cenę wersji podstawowej modelu samochodu w tys. PLN. Przyjąć, że 1 mila to 1.6 km; 1 US gallon to 3.8 litra; 1 funt to 0.4536 kg; 1 $ to 3.35 PLN.

(b) Wyznaczyć podstawowe statystyki próbkowe dla danych opisujących cenę wersji podstawowej samochodu. Obliczyć kwantyl rzędu 0.95 dla tych danych. Wypisać ceny wersji podstawowej samochodów, które były wyższe od wyznaczonego kwantyla. Jakich modeli te ceny dotyczą?

(e) Narysować wykres słupkowy i kołowy dla zmiennej `Type`. Ile, spośród badanych samochodów, zaliczono do kategorii sportowe?

(f) Sporządzić i opisać wykresy skrzynkowe dla zużycia benzyny podczas jazdy w mieście osobno dla samochodów amerykańskich i nieamerykańskich. Wyciągnąć wnioski.

(g) Sporządzić wykres rozrzutu ceny podstawowej wersji samochodu od jego zużycia benzyny w mieście oraz wykres rozrzutu zużycia benzyny w mieście w funkcji zużycia benzyny na autostradzie. Umieścić oba te wykresy w jednym oknie. Wykresy uzupełnić odpowiednimi współczynnikami korelacji.

(h) Narysować histogram częstości dla danych dotyczących wagi samochodu.

Zad. 2 (2 pkt) Zbiór `airpollution.txt` zawiera dane dotyczące związku pomiędzy zanieczyszczeniem powietrza i śmiertelnością w 60 miastach amerykańskich. Wykorzystywane zmienne:

- `Mortality` - skorygowana wiekiem liczba zgonów na 100 000 mieszkańców
- `Education` - mediana liczby lat kształcenia
- `NonWhite` - procent tej podpopulacji
- `income` - mediana zarobków w tys. dolarów
- `JanTemp`, `JulTemp` - średnie temperatury w styczniu i lipcu (w stopniach Fahrenheita)
- `NOx` - stężenie tlenku azotanu

(a) Wczytać dane i dokonać statystycznej analizy wykorzystywanych zmiennych. Dopasować model liniowy ze zmienną objaśnianą `Mortality` i zmienną objaśniającą `NOx`.

(b) Podać współczynnik nachylenia prostej regresji oraz jego błąd standardowy. Sprawdzić czy dopasowany model dobrze opisuje dane.

(c) Dopasować model liniowy ze zmienną objaśnianą `Mortality` i zmienną objaśniającą `log(NOx)`. Podać współczynnik nachylenia prostej regresji oraz jego błąd standardowy. Czy model ten dobrze opisuje dane?

(d) W modelu liniowym ze ze zmienną objaśnianą `Mortality` i zmienną objaśniającą `log(NOx)` znaleźć obserwacje o dużych residuach studentyzowanych. Sporządzić nowy model pomijając owe obserwacje. Porównać wartości współczynnika R^2 dla tych dwóch modeli.

Zad. 3 (3 pkt) Dane w pliku `savings.txt` zawierają informacje dotyczące sytuacji ekonomicznej mieszkańców 50 krajów. Dane są wielkości uśrednione za lata 1960 - 1970:

- `Country` - nazwa kraju
- `Savings` - łączne oszczędności przypadające na osobę podzielone przez dochód netto
- `pop15` - procent populacji poniżej 15 roku życia
- `pop75` - procent populacji powyżej 75 roku życia
- `dpi` - dochód netto przypadający na jednego mieszkańca
- `ddpi` - tempo wzrostu dochodu (w %)

(a) Wczytać dane i dokonać statystycznej analizy wykorzystywanych zmiennych. Dopasować model liniowy opisujący zależność `Savings` od `dpi`, `ddpi`, `Pop15` i `Pop75` .

(b) Narysować wykres reszt w tym modelu. Zidentyfikować, którym krajom odpowiada najmniejsza i największa wartość reszt.

(c) Odczytać i narysować wartości dźwigni (leverage). Dla których krajów wartość dźwigni są duże? Wyznaczyć reszty studentyzowane. Które z nich są duże?

(d) Wyznaczyć wartości miar DFFITS, DFBETAS. Zinterpretuj uzyskane wyniki. Wyznaczyć odległości Cooke'a. Które z nich można uważać za duże? Wskazać wszystkie obserwacje wpływowe

(e) Przeprowadź regresję dla danych z wyłączoną obserwacją o największej wartości odległości Cooke'a. Porównaj nowy model z poprzednio rozważanym. Czy zmienna `dpi` jest istotna w modelu dopasowanym w punkcie (a)?

(f) Narysować wykres zmian wartości współczynników przy zmiennych `pop15` oraz `pop75` w modelu z usuniętą obserwacją. Który kraj ma największy wpływ?

Zad. 4 (2 pkt) Wykorzystując dane zawarte w zbiorze `realest.txt` zbadać zależność ceny domu na przedmieściach Chicago (`Price`) od liczby sypialni (`Bedroom`), powierzchni w stopach kwadratowych (`Space`), liczby pokoi (`Room`), szerokości frontu działki w stopach (`Lot`), rocznego podatku od nieruchomości (`Tax`), liczby łazienek (`Bathroom`), liczby miejsc parkingowych w garażu (`Garage`) i stanu domu (`Condition`, 0-dobry, 1-wymaga remontu).

(a) Dopasować liniowy model regresji opisujący badaną zależność. Jeśli wszystkie pozostałe zmienne objaśniające są ustalone, jaki wpływ na cenę ma zwiększenie liczby sypialni o 1? Znaleźć uzasadnienie tego pozornie błędnego wyniku. Porównać ten wynik z wynikiem otrzymanym dla modelu linowego opisującego zależność ceny domu jedynie od liczby sypialni.

(b) Założyć, że posiadamy dom w tej okolicy, w dobrym stanie, z 3 sypialniami, o powierzchni 1500 stóp kwadratowych, z 8 pokojami, 40 stopami szerokości działki, 5 łazienkami, 1 miejscem w garażu i podatkiem w wysokości 1000 dolarów. Za ile spodziewamy się go sprzedać?

Zad. 5 (2 pkt) Dane w pliku `gala_data.txt` zawierają informacje dotyczące 30 Wysp Galapagos:

- `Species` - liczba gatunków żółwi na danej wyspie
- `Endemics` - liczba gatunków endemicznych
- `Area` - powierzchnia wyspy (w km^2)
- `Elevation` - najwyższe wzniesienie na wyspie (w m)
- `Nearest` - odległość od najbliższej wyspy (w km)
- `Scruz` - odległość od wyspy Santa Cruz (w km),
- `Adjacent` - powierzchnia najbliższej wyspy (w km^2).

(a) Dopasować model liniowy opisujący zależność liczby gatunków żółwi na wyspie od powierzchni wyspy, jej najwyższego wzniesienia, odległości od najbliższej wyspy, odległości od wyspy `Santa Cruz` i powierzchni najbliższej wyspy. Przeprowadzić diagnostykę modelu. Zweryfikować czy wariancja residuów zależy od wartości prognozowanych.

(b) Aby usunąć problem zmiennej wariancji residuów należy spierwiastkować zmienną objaśnianą i dopasować model liniowy z tak przekształconą zmienną. Przeprowadzić diagnostykę nowego modelu. Znaleźć zmienną objaśniającą o największym `p-value` testu weryfikującego hipotezę o istotności poszczególnych zmiennych. Usunąć ją z modelu i dopasować mniejszy model. Porównać wartości współczynników determinacji i zmodyfikowanych współczynników determinacji modeli z punktów (a) i (b).

Zad. 6 (2 pkt) Zbudować drzewo decyzyjne dla danych w pliku `irys.txt` (`http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`), które zawierają informacje dotyczące 150 kwiatów irysów, które opisano 4 cechami (atrybuty warunkowe): długość i szerokość płatka, oraz długość i szerokość łodygi. Dodatkowo mamy atrybut decyzyjny (class) który przyjmuje 3 możliwe wartości: „Iris-setosa", „Iris-Versicolor" oraz „Iris-virginica", które równo dzielą zbiór po 50 obserwacji dla każdej z tych klas (33.3%).
Chcąc wyręczyć botaników w rozpoznawaniu kwiatów, należy zbudować drzewo klasyfikacyjne, które na podstawie 4 parametrów numerycznych kwiatu odgadnie jego gatunek.

(a) Wczytać cały zbiór obserwacji. Podziel dane na zbiór treningowy i testowy. Wyświetlić jak wygląda wytrenowane drzewo w formie tekstowej i graficznej. Wyjaśnić otrzymane reguły.

(d) Wyświetlić tzw. macierz błędów. Wyjaśnić ile błędów popełnił twój algorytm na zbiorze testowym. Jakie gatunki pomylił z jakimi? Podać ile mamy procent dobrze odgadniętych odpowiedzi.

Zad. 7 (1 pkt) Korzystając z danych w pliku `irys.txt` zbudować klasyfikator k-NN, który przydziela rekordowi (np. irysowi) odpowiednią klasę (czyli w przypadku irysów: gatunek) szukając k najbardziej podobnych instancji do niego i przydzielając mu taką klasę, jaką ma większość z tych rekordów.

(a) Wczytaj bazę danych irysów. Znormalizować dane liczbowe i zapisz ją pod `iris.norm`. Podzielić iris.norm na zbiór treningowy i testowy. Uruchomić algorytm 3-najbliższych sąsiadów dla zbioru treningowego i testowego. Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?

(c) Dokonać ewaluacji klasyfikatora i wyświetl macierz błędu, oraz jego dokładność.

Pliki z rozwiązaniami (DM_Jan_Kowalski.pdf) oraz funkcjami (pliki R) należy przesłać jako niespakowane załączniki jednym listem elektronicznym o temacie **DM PROJEKT IND** na adres:

zofia.grudziak@pw.edu.pl

W treści listu należy podać swoje imię i nazwisko.