Podstawy Przetwarzania Danych

Laboratoria 6: Selekcja zmiennych odstające

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

Wersja 1.0
5 marca 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

F - Test

```
var.test(x, y, ratio = 1,
         alternative = c("two.sided", "less", "greater"
             ),
         conf.level = 0.95)
```

- Funkcja var.test porównuje wariancje dwóch populacji x i y o rozkładzie normalnym.
- Hipotezą zerową jest założenie, że stosunek wariancji wynosi ratio, domyślnie 1.
- Hipoteza alternatywna jest definiowana przez parametr alternative.
  - Możemy założyć, że stosunek jest inny niż dla hipotezy zerowej lub, że jest mniejszy lub większy.
- Poziom ufności testu ustalamy parametrem conf. level domyślnie 0.95.

Analiza wariancji

```
1 aov.res<-aov(formula, data)
```

- Funkcja aov przeprowadza analizę wariancji na danych opisanych zmienną data.
- Zmienna formula przedstawia formułę postaci *obserwowane_zmienne* $\sim$ *zmienna_kategoryzujca*

Wyniki analizy

```
1 summary(aov.res)
```

- Funkcja summary zwraca opis uzyskanych wyników analizy.
- W szczególności można uzyskać poziom istotności uzyskanych wyników.

Test Tukeya

```
1 TukeyHSD(aov.res)
```

- Test Tukeya pozwala nam lepiej zbadać zróżnicowanie zmiennej w poszczególnych klasach
- Funkcja TukeyHSD bazuje na wynikach analizy, aby pokazać zależności między klasami.

Zadanie 1

- Dla danych z pliku fauna.csv porównaj wariancję zmiennej ciężar_max z pozostałymi zmiennymi.
- Sprawdź, czy prawdopodobieństwo testowe pozwala na odrzucenie hipotezy zerowej i przyjęcie, że któraś ze zmiennych ma wariancję różną od ciężar_max.

Zadanie 2

- Dla danych z pliku fauna.csv porównaj wariancję zmiennej ciężar_max z pozostałymi zmiennymi.
- Sprawdź, czy prawdopodobieństwo testowe pozwala na odrzucenie hipotezy zerowej i przyjęcie, że któraś ze zmiennych ma wariancję mniejszą od ciężar_max.
- Sprawdź, czy prawdopodobieństwo testowe pozwala na odrzucenie hipotezy zerowej i przyjęcie, że któraś ze zmiennych ma wariancję większą od ciężar_max.
- Czy zmiana hipotezy alternatywnej zmienia uzyskane wyniki?

Zadanie 3

- Dla danych z pliku fauna.csv przeprowadź dla zmiennej Rodzina analizę wariancji.
- Sprawdź czy zmienna ciężar_max jest istotna przy klasyfikacji na rodziny.
- Sprawdź, które klasy są dobrze separowane przez zmienną.
- Czy zmienna ogon_min będzie lepszym czy gorszym wyborem niż zmienna ciężar_max?

Regresja LASSO

```
1 library(glmnet)
2 reg.lasso<-glmnet(x, y, alpha = 1)
```

- Regresja LASSO jest implementowana przez funkcję glmnet.
- Należy określić zmienne objaśniające x i zmienną objaśnianą y.
- Zmienne powinny być przekazane jako macierz. Dane można zrzutować do postaci macierzy używając funkcji as.matrix.

Elementy obiektu LASSO

- Z obiektu regresji LASSO można odczytać następujące elementy

df	liczba zmiennych z niezerowymi parametrami β
lambda	wartości współczynników λ
beta	wartości współczynników β

Ścieżki LASSO

```
plot(reg.lasso, xvar = c("norm", "lambda", "dev"),
     label=FALSE)
```

- Funkcja plot pozwala wykreślić ścieżki LASSO.
- Zmienna xvar określa współczynnik względem którego wykreślamy ścieżki

  norm względem normy L_1
  lambda względem wartości λ
  dev względem wytłumaczonej zmienności.
- Zmienna label pozwala dodać dodatkowe etykiety opisujące zmienne.

Predykcja

```
1 library(glmnet)
2 predictions<-predict(reg.lasso,data)
```

- Utworzony model można wykorzystać do predykcji, jak dowolny model regresyjny.
- Służy do tego funkcja predict

Zadanie 4

- Dla danych z pliku fauna.csv zbuduj model regresji LASSO objaśniając ciężar_max pozostałymi zmiennymi.
- Wykreśl ścieżki LASSO dla utworzonego modelu.
- Określ, która zmienna zostanie wyeliminowana jako pierwsza.

Drzewo decyzyjne

```
1 library(rpart)
2 decision.tree<-rpart(formula,data)
```

- Drzewa decyzyjne są tworzone przez funkcję rpart.
- W podstawowej postaci funkcja buduje drzewo na podstawie zbioru danych data i formuły formula określającej zależności między zmienną objaśnianą, a zmiennymi objaśniającymi.

Wykreślanie drzewa

```
1 library(rpart.plot)
2 rpart.plot(decision.tree)
```

- Pakiet rpart.plot oferuje narzędzia do kreślenia drzewa.
- Są one bardziej eleganckie niż schematy kreślone funkcją plot.

Szczegółowe informacje o drzewie

```
1 summary(decision.tree)
```

- Polecenie summary przedstawia szczegółowy opis drzewa.
- W szczególności pozwala ocenić istotność zmiennych.

Zadanie 5

- Dla danych z pliku fauna.csv zbuduj model klasyfikacyjny drzewa decyzyjnego przypisujący obserwacje do Rodzin.
- Wykreśl drzewo decyzyjne. Określ na podstawie drzewa, które zmienne będą istotne dla budowanego modelu.
- Zweryfikuj obserwację na podstawie istotności cech zwracanych przez algorytm budowy drzewa.
- Która zmienna jest najbardziej a która najmniej istotna?

Zadanie 6

- Dla danych z pliku fauna.csv zbuduj model klasyfikacyjny drzewa decyzyjnego przypisujący obserwacje do Rodzin.
- Z budowanego modelu usuń zmienne, które były użyte do budowy poprzedniego drzewa.
- Zbadaj istotność użytych zmiennych. Jak zmieniła się ona w porównaniu do poprzedniego drzewa?