Statystyczna analiza danych
Laboratorium nr 6

30.05.2023

Spis treści

1 Zastosowanie ANOVA - Wybór cech

ANOVA można zastosować do wyboru cech, badając związek między zmiennymi predykcyjnymi (cechami) a zmienną docelową. Obliczając statystykę F i powiązaną wartość p dla każdej cechy, można określić, czy istnieje statystycznie istotna zależność między cechą a zmienną przewidywaną.

W kontekście wyboru cech, ANOVA jest zwykle używana do wybierania cech numerycznych, gdy zmienna przewidywana jest kategoryczna. Pomaga zidentyfikować cechy, które mają znaczący wpływ na zmienną docelową i może być wykorzystana do odfiltrowania cech nieistotnych lub nieinformacyjnych.

Ogólna procedura korzystania z ANOVA do wyboru cech obejmuje następujące kroki:

1. Oblicz statystykę ANOVA (statystyka F) i powiązaną wartość p dla każdej cechy.
2. Ustaw poziom istotności (np. 0,05), aby określić próg istotności statystycznej.
3. Wybierz cechy, które mają wartości p poniżej wybranego poziomu istotności.
4. Użyj wybranych cech do dalszej analizy lub budowy modelu.

Należy zauważyć, że ANOVA zakłada pewne założenia dotyczące danych, takie jak normalność i jednorodność wariancji. Jeśli te założenia zostaną naruszone, konieczne mogą być alternatywne metody lub transformacje. Ponadto ANOVA jest bardziej odpowiednia dla liniowych zależności między cechami a zmienną docelową, więc może nie być najlepszym wyborem dla złożonych lub nieliniowych zależności. W takich przypadkach bardziej odpowiednie mogą być inne metody selekcji cech, takie jak wzajemna informacja lub rekurencyjna eliminacja cech.

1.1 Problem wielokrotnych porównań

Problem porównań wielokrotnych – w statystyce zjawisko występujące przy dokonywaniu estymacji lub weryfikacji hipotez statystycznych polegające na zwiększonym ponad nominalny poziom istotności ryzyku omyłkowego przyjęcia fałszywej hipotezy alternatywnej (popełnienia błędu I rodzaju) przy wykonywaniu wielu porównań tej samej grupy (rodziny) hipotez jednocześnie. Przynajmniej jeden z testów może przypadkiem, dzięki losowej zmienności prób, przekroczyć próg istotności.

Przykładowo, choć w rzeczywistości w populacji badane zjawisko nie występuje w żadnym stopniu, badacz, który wykona kilkaset porównań bez odpowiedniej poprawki w podgrupach według płci, wieku, wykształcenia, klasy socjoekonomicznej, miejsca zamieszkania – np. w modelu 2 płcie $\times$ 5 grup wiekowych $\times$ 5 grup wykształcenia $\times$ 3 klasy socjoekonomiczne $\times$ 3 typy miejsca zamieszkania, co daje 450 porównań – znajdzie praktycznie na pewno bardzo wiele przypadkowo istotnych statystycznie różnic. Nawet jeśli badane zjawisko rzeczywiście istnieje, zaburzona kontrola błędu I rodzaju powoduje przeszacowywanie jego wielkości efektu.

Jedną z metod kontroli błędu I rodzaju w rodzinie testów jest procedura Benjaminiego-Hochberga.

1.2 Przykład

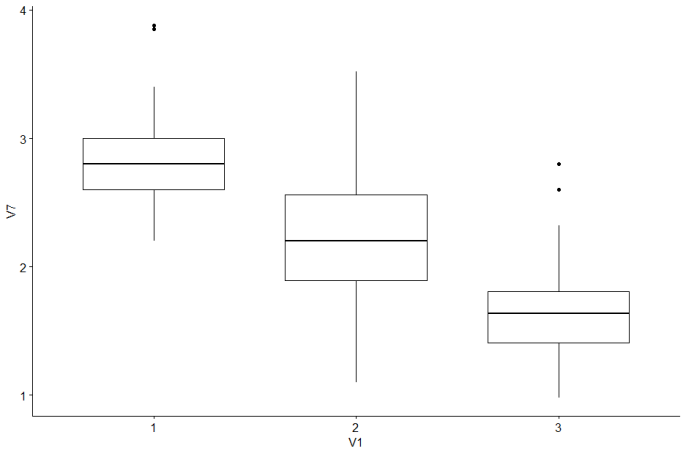
Załóżmy, że zmienna `data` zawiera dane o `n` cechach (kolumnach) i `m` obserwacjach (wierszach) oraz, że zmienna zależna (odpowiedzi) nazywa się `target` (i stanowi "n plus pierwszą" kolumnę w ramce).

```
# Liczba cech - odejmujemy ostatnia kolumne ze zmienna odpowiedzi
n <- ncol(data) - 1

# Stworz puste wektory do przechowywania p-wartosci oraz skorygowanej p-wartosci
p_values <- vector(mode = "numeric", length=n)
adjusted_p_values <- vector(mode = "numeric", length=n)

# Przeprowadz testy ANOVA oraz oblicz p-wartosc dla kazdej cechy
for (i in 1:n) {
    anova_result <- anova(lm(data[,i] ~ data[,target]))
    p_values[i] <- anova_result$`Pr(>F)`[1]  # wyciagniecie p-wartosci z tablicy ANOVA
}

# Korekcja metoda Benjamini-Hochberg'a
adjusted_p_values <- p.adjust(p_values, method="BH")

# Wyswietlenie wynikow
for (i in 1:length(adjusted_p_values)) {
    cat("Feature ", names(data)[i], ": P-value = ", p_values[i], ", Adjusted p-value = ", adjusted_p_values[i], "\n")
}
```

`summary(aov(...))` oraz `anova(lm(...))` dają tożsame wyniki, ale na wynikach tej drugiej konstrukcji łatwiej operować.

1.3 Porównanie kilku grup na jednym wykresie

```
> library("ggpubr")
> ggboxplot(wine, x="V1", y="V7")
```



Zadanie nr 1

- Załóżmy, że zbiór `iris` spełnia założenia testu ANOVA.
- Za pomocą testu ANOVA zidentyfikuj, które ze zmiennych byłyby dobrymi predyktorami w modelu klasyfikacyjnym oraz w jakiej kolejności (im mniejsza p-wartość, tym lepszy predyktor). Przyjmij pozom istotności $\alpha = 0.05$.
- Pamiętaj o korekcji błędu I rodzaju.
- Dla każdego z wybranych predyktorów stwórz wykres pudełkowy porównujący rozkłady zmiennej w grupach.

2 Test Kruskala-Wallisa

Istnieje nieparametryczna wersja testu ANOVA nazywająca się testem Kruskala-Wallisa. Test Kruskala-Wallisa jest używany gdy założenia normalności lub jednorodności wariancji nie są spełnione lub gdy dane są porządkowe lub rangowe. Test Kruskala-Wallisa weryfikuje istnienie istotnych różnic pomiędzy medianami dwóch lub więcej niezależnych grup.

- Hipoteza zerowa H_0: Nie ma istotnych różnic pomiędzy medianami w grupach.
- Hipoteza alternatywna H_a: Mediana co najmniej jednej grupy istotnie różni się od pozostałych.
- Interpretacja: Jeżeli wartość p jest mniejsza niż poziom istotności, możemy stwierdzić, że istnieją istotne różnice między grupami.

Test można wykonać za pomocą funkcji `kruskal.test()` w następujący sposób:
`kruskal.test(cecha ~ grupa, dane = moje_dane)`

Zadanie nr 2

- Pobierz zbiór do klasyfikacji jakości win `winequality.csv`. Ostatnia kolumna odpowiada jakości wina.
- Zidentyfikuj za pomocą właściwego testu, które z tych zmiennych byłyby dobrymi predyktorami w modelu klasyfikacyjnym na poziomie istotności $\alpha = 0.05$.
  - Pamiętaj o sprawdzeniu założeń.
  - Pamiętaj o korekcji błędu I rodzaju.