IRD kolokwium 2020/2021, semestr zimowy, wariant C 27.01.2021

Formalności

Rozwiązania wpisuj w pliku o nazwie utworzonej wedle wzorca: "Nazwisko_NrIndeksu.R". Nie musisz przeklejać poleceń, ale zachowaj kolejność zadań i wypisuj wyraźnie ich numery (jako komentarze). Zachowanie struktury pliku przyspieszy sprawdzanie i ogłoszenie wyników.

Przed końcem kolokwium wyślij plik w zakładce Zadania aplikacji Teams. Dla pewności, dodatkowo, możesz również wysłać kopię rozwiązań **jako załącznik** do wiadomości mailowej na adres **nosarzewski.aleks@gmail.com**. Tytuł maila powinien zawierać Twój numer indeksu.

Kolokwium zostanie sprawdzone przez uruchamianie kodu linijka po linijce. Upewnij się, że fragmenty kodu są we właściwej kolejności oraz cały skrypt wykonuje się poprawnie.

Zadanie 1 (10 pkt)

Napisz funkcję `twins`, która dla dwóch liczb `n` i `m` zwróci `TRUE`, jeżeli liczby te są liczbami bliźniaczymi oraz `FALSE` w przeciwnym przypadku. Możesz korzystać wyłącznie z funkcjonalności dostępnych w podstawowej instalacji R.

Dwie liczby są liczbami bliźniaczymi, jeżeli obie są liczbami pierwszymi, zaś różnica między nimi wynosi 2 (link).

PODPOWIEDŹ: Do napisania tej funkcji możesz napisać pomocniczą funkcję, która będzie sprawdzała czy dana liczba jest liczbą pierwszą, a następnie zastosować ją w ciele funkcji `twins()`. Przykładowe algorytmy sprawdzające czy dana liczba jest pierwsza możesz znaleźć tutaj (link).

Twoja funkcja powinna działać w następujący sposób:

```
> twins(199, 197)
TRUE
> twins(2, 5)
FALSE
> twins(4, 6)
FALSE
```

Zaaplikuj funkcję na dowolnych dwóch liczbach będącymi liczbami bliźniaczymi.

Zadanie 2 (10 pkt)

Podpowiedź do zadania - do danych z pakietu `caret` można dostać się w następujący sposób:

```
library("caret")
data("GermanCredit")
```

Na podstawie danych `GermanCredit` z pakietu `caret`:

- Z wykorzystaniem operatora pipeline (`%>%`) z pakietu `dplyr`, oblicz następujące statystyki dla wieku (*Age*): minimum, średnią, medianę oraz maksimum w zależności od długości zamieszkania w danym miejscu(*ResidenceDuration*). Wyniki posortuj malejąco według wartości zmiennej grupującej. Jaka długość charakteryzowała się najwyższym średnim wiekiem? (odpowiedź napisz jako komentarz)
- Narysuj wykres (wykorzystując pakiet `ggplot2`) taki sam jak w Załączniku 1.

Zadanie 3 (20 pkt)

Podpowiedź do zadania - do danych z pakietu `geepack` można dostać się w następujący sposób:

```
#install.packages("geepack")
library("geepack")
data("seizure")
```

Na podstawie danych `seizure` z pakietu `geepack`:

a) Wczytaj dane, przekonwertuj zmienną `trt` na typ czynnikowy, ustaw ziarno losowe na swój numer indeksu, oraz podziel zbiór na uczący i testowy w proporcji 70% do 30%.
b) Zbuduj dwa modele klasyfikujące zmienną `trt`:
   - drzewo klasyfikacyjne z wszystkimi zmiennymi objaśniającymi (ustaw maksymalną głębokość drzewa na 4),
   - las losowy z wszystkimi zmiennymi objaśniającymi (ustaw liczbę wykorzystanych drzew na 300).
c) Która ze zmiennych objaśniających dla modelu lasu losowego ma najmniejszy wpływ na zmienną prognozowaną (odpowiedź napisz jako komentarz)? Nie musisz analizować co reprezentuje ta zmienna.
d) Na podstawie zbioru testowego, policz `F1` oraz `Negative Predictive Value`. Oceń, który model lepiej prognozuje zmienną `trt` ze względu na `Negative Predictive Value`? (odpowiedź napisz jako komentarz)
e) Dla lepszego modelu (z podpunktu d)) policz AUC (napisz jego wartość jako komentarz) oraz narysuj ROC. Nie przejmuj się jeżeli wykres będzie mieć strukturę schodkową.

Zadanie 4 (10p)

Podpowiedź do zadania - do danych z pakietu `datasets` (pakiet wchodzi w skład podstawowej instalacji R - nie musisz go instalować ręcznie) można dostać się w następujący sposób:

```
data("esoph")
```

Na podstawie danych `esoph`:

a) Podziel zbiór na uczący i testowy w proporcji 80% do 20% (ziarno losowe ustaw na swój numer indeksu).
b) Na zbiorze uczącym zbuduj modele prognozującą zmienną `ncases`:
   - regresji liniowej z wszystkimi zmiennymi jako zmiennymi objaśniającymi,
   - drzewa regresyjnego z wszystkimi zmiennymi jako zmiennymi objaśniającymi (ustaw wartość *complexity parameter* na 0.08).
c) Podaj słownie jedną przykładową regułę otrzymaną z drzewa jako komentarz, pamiętaj o różnicy w znaczeniu prognozy między drzewem klasyfikacyjnym a regresyjnym. (Nie musisz analizować znaczenia zmiennych, wystarczy podać konkretną nazwę i warunek.)
d) Na podstawie zbioru testowego oblicz:
   - błąd średniokwadratowy (MSE),
   - względny błąd absolutny (RAE).
e) Który model jest lepszy (odpowiedź napisz jako komentarz i uzasadnij)?

Załącznik 1


Wartosc kredytow w zaleznosci od ich klasy, w podziale
na pracownikow krajowych oraz zagranicznych (ForeignWorker)
0
1
Wartosc kredytu
15000
10000
5000
0
Bad
Good
Bad
Good
Klasa kredytu