Podstawy Przetwarzania Danych

Laboratoria 13: Ocena estymatorów

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.

Wczytywanie danych

```
require(datasets)
data(iris)
summary(iris)
```

- R pozwala na wczytywanie predefiniowanych zbiorów danych z repozytorium. Służy do tego pakiet dataset.
- Zbiory wczytujemy poleceniem data.
- Zawartość zbioru można podsumować poleceniem summary.

Zadanie 1

- Wczytaj zbiór iris i usuń z niego klasę setosa.
- Następnie, korzystając z 10-krotnej kroswalidacji, wylicz wyniki predykcji Species dla metod
  - Native Bayes
  - Multi Layer Perceptron
  - Decision Tree (rpart)
  - Random Forest

Macierz pomyłek

```
1 confusionMatrix(predictions,data)
```

- Polecenie confusionMatrix wylicza macierz pomyłek na podstawie predykcji i docelowych wartości.
- Macierz pomyłek zawiera liczbę elementów z każdej klasy, przypisanej do każdej z klas.
- W przypadku zadania binarnego macierz pomyłek przybiera formę

	Condition P	Condition N
Predicted P	T(rue)P(ositive)	F(alse)P(ositive)
Predicted N	F(alse)N(egative)	T(rue)N(egative)

Wyliczanie ROC

```
require(pROC)
ROC <- roc(response, prediction)
```

- Funkcja roc z pakietu pROC wylicza ROC.
- Jako parametry podajemy:
  - wektor klas do których należą obserwacje response,
  - wektor predykcji dla obserwacji prediction .

Kreślenie ROC

```
require(ggplot2)
ggroc(roc)
ggroc(list(roc, roc1))
```

- Funkcja ggroc z pakietu ggplot2 kreśli wykres ROC.
- Funkcja pozwala na wykreślenie jednej krzywej ROC lub wielu przekazywanych jako lista.

Zadanie 2

- Porównaj wyniki uzyskane przez poszczególne metody stosując
  - Skuteczność,
  - F-measure,
  - AUC.

Analiza statystyczna wyników

```
1 t.test(acc1,acc2)
```

- Mając do dyspozycji wyniki kilku niezależnych testów danej metody możemy powiedzieć czy uzyskane przez nie wyniki różnią się istotnie statystycznie.
- Test t Studenta pozwala sprawdzić czy rozkłady wyników dwóch metod różnią się od siebie.
- Test t Studenta jest zaimplementowany funkcją `t.test`.

Zadanie 3

- Odczytaj wyniki skuteczności z 10 testów kroswalidacji dla klasyfikatorów dających najniższą i najwyższą skuteczność.
- Sprawdź, czy różnice między ich skutecznością są statystycznie istotne.
- Przyjmij poziom istotności 0.05.

Test Wilcoxona

```
wilcox.test(x, y, paired = TRUE,
  alternative = c("two.sided", "less", "greater"))
```

- Funkcja wilcox.test implementuje test Wilcoxona.
- Parametr paired określa, iż dane w wektorach x i y są sparowane.
- Parametr alternative określa hipotezę alternatywną.
  - rozkłady są równe,
  - pierwszy rozkład ma mniejszą średnią,
  - pierwszy rozkład ma większą średnią

Zadanie 4

- Wykonaj 10-krotną kroswalidację dla lasu losowego i regresji grzbietowej, aby wyliczyć Sepal.Length.
- Wylicz błędy uzyskane dla obu metod.
- Przeprowadź testy statystyczne, aby sprawdzić czy na poziomie istotności 0.05
  - Rozkłady błędów są różne,
  - Jeden model regresji daje mniejsze błędy.