Podstawy Przetwarzania Danych

Laboratoria 11: Testowanie

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.

Środowisko testowe

```
1 train.control <- trainControl(method = "LOOCV")
2 train.controlCV <- trainControl(method = "cv", number
      =10)
```

- Metoda trainControl pozwala określić sposób testowania algorytmów uczenia maszynowego.
- Parametr method określa sposób estymacji jakości algorytmu.
- Niektóre metody estymacji wymagają określenia dodatkowych parametrów jak wartość number dla algorytmu kroswalidacji.

Metody

- Przykładowe metody testowania

boot	Estymator błędu prób bootstrapowych
boot632	Estymator .632
cv	krosswalidacja
LOOCV	Leave-one-out cross-validation
none	cały zbiór
oob	Out-of-bag

Uruchomienie testu

```
model <- train(formula, data, method,
                        trControl = train.control)
```

- Metoda train pozwala uruchomić zdefiniowane testy dla danej metody uczenia maszynowego.
- Parametr method określa algorytm uczenia maszynowego.
- Zadanie do rozwiązania jest zdefiniowane przez parametry data i formula.
- Informacje o sposobie testowania przekazujemy za pomocą parametru trControl.
- W wyniku działania metody powstaje model, który możemy użyć do dalszej pracy.

Metody uczenia maszynowego

- Przykładowe obsługiwane metody
  - Bagging
  - Bayesian Methods
  - Boosted Trees
  - Elastic Net
  - K Nearest Neighbor
  - Neural Networks
  - Random Forests
  - Support Vector Machines
- Metody mogą wymagać dodatkowych parametrów wywołania.

Wykorzystanie modelu

```
1 predict(model, data)
```

- Metoda train zwraca najlepszy z modeli powstałych podczas testowania.
- Model może być wykorzystywany do predykcji, używając metody predict

Ocena wyników regresji

R-squared R^2 reprezentuje kwadrat korelacji między predykcjami, a oczekiwanymi wynikami.

$$R^2 = 1 - \frac{\sum_1^n (o_i - p_i)^2}{\sum_1^n (o_i - \bar{o}_i)^2}$$

Root Mean Squared Error błąd średnio-kwadratowy między predykcjami, a oczekiwanymi wynikami.

$$RMSE = \frac{1}{n}\sqrt{\sum_1^n (o_i - p_i)^2}$$

Mean Absolute Error średni błąd bezwzględny między predykcjami, a oczekiwanymi wynikami.

$$MAE = \frac{1}{n}\sum_1^n \|o_i - p_i\|$$

Zadanie 1

- Dla danych z pliku mfeb15.csv oszacuj średnią prędkości AverageSpeed używając zmiennych Date,TimePeriod i LinkLength
- Do zbioru uczącego wybierz rekordy isLearning =TRUE.
- Zbuduj referencyjny model testujący nie stosując żadnych specjalnych metod estymacji.
- Jako algorytm uczenia maszynowego wybierz drzewo decyzyjne rpart .
- Przetestuj wyniki modelu dla zbioru testowego isLearning =FALSE

Zadanie 2

- Powtórz poprzednie zadanie stosując metody

boot	Estymator błędu prób bootstrapowych
boot632	Estymator .632
cv	krosswalidacja
LOOCV	Leave-one-out cross-validation

- Porównaj uzyskane wyniki ze zbiorem referencyjnym

Zadanie 3

- Powtórz poprzednie zadania dla lasów losowych.
- Porównaj otrzymane wyniki, aby odpowiedzieć na pytania
  - Która metoda daje lepsze wyniki (drzewo czy las)?
  - Który sposób uczenia daje lepsze wyniki?