WSTĘP DO ANALIZY DANYCH

Lab. 5: Wykorzystanie drzew do zadań regresji - drzewa regresyjne i drzewa modeli

v. 1.0.0_bodp

PLAN LAB. 5

1. Zasada działania drzew regresyjnych i drzew modeli.
2. Algorytmy CART i Cubist.
3. Analiza danych rzeczywistych: budowa modelu oceny jakości produktu spożywczego na przykładzie przemysłu winiarskiego.

DRZEWA DO REGRESJI

- Odpowiednio zmodyfikowane drzewa decyzyjne można wykorzystać również do zadań regresji.
- Drzewa do predykcji zmiennych ilościowych dzielą się w zasadzie na dwa typy:

  drzewa regresyjne (wprowadzone jako część przełomowego algorytmu **CART** — *classification and regression tree*) — predykcja na podstawie średniej wartości przykładów z liścia;

  drzewa modeli (wprowadzone później niż drzewa regresyjne, są mniej znane, ale często skuteczniejsze) — budowane mniej więcej w ten sam sposób co drzewa regresyjne, natomiast dla każdego liścia estymowany jest model regresji wieloczynnikowej na podstawie przykładów, które w tym liściu się znajdują.

- Liczba modeli regresji wieloczynnikowej w drzewie modeli równa się liczbie liści w tym drzewie. Może się więc zdarzyć, że takich modeli regresji budowanych są dziesiątki, a nawet setki. To sprawia, że model główny jest trudniejszy w interpretacji w porównaniu do odpowiadającego mu drzewa regresyjnego. Jest to cena, którą płacimy za często większą dokładność drzewa modeli.

DRZEWA REGRESYJNE I DRZEWA MODELI

- mocne i słabe strony (ogólnie oraz w porównaniu do regresji liniowej)

Mocne strony	Słabe strony
Wykorzystanie mocnych stron drzew decyzyjnych do zadań regresji	Nie są tak dobrze znane jak regresja liniowa
Automatyczny wybór cech - przydatne przy zbiorach danych o bardzo dużym wymiarze (liczbie cech)	Wymagają dużej liczby przykładów w zbiorze uczącym
Dają możliwość modelowania innych niż tylko liniowe struktur w danych (przewaga w przypadku zadań z wielo-ma cechami lub wieloma złożonymi nieliniowymi zależnościami pomiędzy cechami a zmienną wynikową)	Trudno jest ocenić wpływ poszczególnych cech na wynik (na pewno trudniej niż w przypadku regresji liniowej)
Modele łatwe w interpretacji	Bardzo duże drzewa mogą być trudniejsze w interpretacji niż model regresji liniowej

BUDOWA DRZEW DO REGRESJI

- Drzewa do regresji budowane są w analogiczny sposób jak drzewa decyzyjne do klasyfikacji. Zaczynając od korzenia, dane dzielone są na kolejne podzbiory zgodnie ze strategią „dziel i zwyciężaj".
- Wybierane są podziały, które skutkują największym wzrostem jednorodności zmiennej objaśnianej w tworzonych podzbiorach danych.
- W przypadku drzew klasyfikacyjnych jednorodność przykładów w liściach mierzyliśmy za pomocą entropii. W przypadku drzew do regresji jednorodność mierzy się najczęściej za pomocą statystyk takich jak wariancja, **odchylenie standardowe** lub odchylenie przeciętne (średnie odchylenie bezwzględne od średniej).

REDUKCJA ODCHYLENIA STANDARD.

- Jednym z często używanych kryteriów wyboru podziału jest **redukcja odchylenia standardowego** (ang. *standard deviation reduction*, SDR):

$$\text{SDR} = sd(T) - \sum_{i}^{n} \frac{|T_i|}{T} sd(T_i),$$

$sd(T)$ — odchylenie standardowe wartości ze zbioru T,

$T_1, T_2, \ldots, T_n$ są podzbiorami zbioru T powstałymi w wyniku danego podziału,

$|T|$ — liczba przykładów w T.

REDUKCJA ODCHYLENIA STANDARD.

Zadanie 1. Rozważmy przykładową sytuację dwóch opcji podziału: względem binarnej zmiennej A i względem binarnej zmiennej B (rysunek).

dane przed podziałem	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
po podziale A	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
po podziale B	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
							T_1						T_2		

Korzystając z kryterium redukcji odchylenia standardowego, wybrać lepszy podział.

PRZYKŁAD NA DANYCH RZECZYWISTYCH

Model oceny jakości produktu spożywczego na przykładzie przemysłu winiarskiego.

- Tego typu modele potencjalnie mogą działać lepiej w porównaniu do testów wykonywanych przez ekspertów: oceny będą bardziej obiektywne, spójne i uczciwe.
- Jest wiele czynników, które mają wpływ na rentowność winnic: pogoda, warunki środowiskowe, technologie produkcji i butelkowania, także sam marketing i cena mogą mieć wpływ na to, jak produkt zostanie odebrany przez klienta.
- Zastosowania uczenia maszynowego w przemyśle winiarskim: identyfikacja kluczowych różnic w składzie chemicznym win z różnych regionów, identyfikacja składników chemicznych, które powodują, że wino ma słodszy smak, ocena jakości wina (trudne zadanie).

PRZYKŁAD NA DANYCH RZECZYWISTYCH

Model oceny jakości produktu spożywczego na przykładzie przemysłu winiarskiego.

- Cel naszej analizy: **zbudowanie modelu eksperckiego**, który będzie w stanie naśladować oceny ekspertów winiarstwa.
- Do budowy modelu wykorzystamy drzewa, ponieważ ich struktura jest prosta w interpretacji i można je wykorzystać do identyfikacji kluczowych czynników, które wpływają na daną (wysoką lub niską) ocenę wina.

KROK 1. ZBIERANIE DANYCH

- Do opracowania modelu wykorzystamy dane *Wine Quality Data Set*: `https://archive.ics.uci.edu/ml/datasets/wine+quality`.
- W zbiorze znajdują się przykłady win białych i czerwonych typu Vinho Verde z Portugalii. Zbudujemy model dla bardziej popularnej odmiany białej.
- Zbiór danych zawiera informacje nt. **11 właściwości chemicznych 4898 przykładów wina białego**, np. kwasowość, zawartość cukru, chlorków, siarki, alkoholu, pH, gęstość.
- **Wszystkie wina były oceniane przez panel ekspercki** (nie mniej niż trzech ekspertów) w ślepej próbie w skali od 0 (bardzo złe) do 10 (doskonałe).
- Dane zostały wcześniej wykorzystane w analizie opisanej w artykule *Modeling wine preferences by data mining from physicochemical properties, Cortez P., Cerdeira A., Almeida F., Matos T., Reis J., Decision Support Systems, 2009, 47, 547–553*.

KROK 2. EKSPLORACJA I PRZYGOTOWANIE DANYCH

Zadanie 2.

a) Wczytać dane do ramki danych w R. Wyświetlić jej strukturę.

b) Zbadać rozkład zmiennej objaśnianej, np. narysować jej histogram. Skomentować otrzymane wyniki.

c) Przeglądnąć podstawowe statystyki dla wszystkich zmiennych w zbiorze danych.

d) Utworzyć zbiór danych uczących i zbiór danych testowych. Do zbioru uczącego wybrać pierwszych 3750 przykładów (około 75%, jak w badaniach Corteza i współautorów — aby porównać wyniki; zbiór danych został wcześniej losowo posortowany).

KROK 3. BUDOWA MODELU

- Zaczniemy od „wytrenowania" drzewa regresyjnego. Do tego celu wykorzystamy funkcję `rpart()` z pakietu **rpart** (*recursive partitioning*).

Składnia metody drzew regresyjnych

z użyciem funkcji `rpart()` z pakietu **rpart**

Budowa modelu:

```
m <- rpart(dv ~ iv, data = mydata)
```

- `dv` to zmienna objaśniana
- `iv` to formuła specyfikująca zmienne objaśniające
- `data` to ramka danych, w której znajdują się zmienne `dv` oraz `iv`

Funkcja zwraca model drzewa regresyjnego, który można wykorzystywać do wykonywania predykcji.

Składnia metody drzew regresyjnych

z użyciem funkcji `rpart()` z pakietu **rpart**

Predykcja:

```
p <- predict(m, test, type = "vector")
```

- `m` to model zbudowany z wykorzystaniem funkcji `rpart()`
- `test` jest ramką danych zawierającą zbiór testowy (z tymi samymi cechami co zbiór uczący)
- `type` określa rodzaj zwracanej wielkości, domyślnie = `"vector"` i zwracane są estymowane wartości zmiennej objaśnianej (średnie dla przykładów z liści)

Funkcja zwraca wektor przewidzianych wielkości (w zależności od ustawienia parametru `type`).

Przykład:

```
wine_model <- rpart(quality ~ alcohol + sulfates, data = wine_train)
wine_prediction <- predict(wine_model, wine_test)
```

KROK 3. BUDOWA MODELU

Zadanie 3.

a) Zbudować model drzewa regresyjnego w rozważanym zadaniu.

b) Wyświetlić podstawowe informacje nt. zbudowanego modelu (wyświetlić wprost utworzony obiekt oraz użyć funkcji `summary()`).

KROK 3. BUDOWA MODELU

Wizualizacja drzew.

- Jednym z pakietów, który umożliwia rysowanie drzew jest pakiet **rpart.plot**.
- Pakiet daje spore możliwości eleganckiego rysowania drzew: `http://www.milbo.org/rpart-plot/`.
- Funkcja `rpart.plot()` umożliwia narysowanie diagramu drzewa na podstawie obiektu typu `rpart`.

Zadanie 4. Narysować drzewo regresyjne zbudowane w poprzednim zadaniu. Przetestować działanie różnych parametrów funkcji `rpart.plot()`.

KROK 4. OCENA MODELU

Zadanie 5.

a) Zastosować zbudowany model drzewa regresyjnego do przykładów win ze zbioru testowego.

b) Wstępnie ocenić jakość predykcji na podstawie porównania podstawowych statystyk dla wyestymowanych wartości oraz dla prawdziwych wartości (np. użyć funkcji `summary()`).

c) Policzyć korelację pomiędzy wartościami wyestymowanymi a prawdziwymi (jest to prosty sposób na wstępne ocenienie jakości zbudowanego modelu).

KROK 4. OCENA MODELU

Średni błąd bezwzględny.

- Jednym z powszechnie stosowanych sposobów kwantyfikowania jakości modeli regresyjnych jest mierzenie **średniej odległości prognoz od prawdziwych wartości**.
- **Średni błąd bezwzględny** (ang. *mean absolute error*, MAE) definiowany jest w następujący sposób:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i|,$$

gdzie n to liczba wykonanych prognoz, natomiast $|e_i|$ to błąd bezwzględny i-tej prognozy.

Zadanie 6. Wyznaczyć średni błąd bezwzględny prognoz wykonanych przez nasz model.

KROK 4. OCENA MODELU

- Średni błąd bezwzględny wynosi około 0.59. Jakość win mierzona jest na skali od 0 do 10. Wynik zatem wydaje się na pierwszy rzut oka całkiem dobry.
- Należy jednak pamiętać, że większość win z naszego zbioru danych nie miała skrajnych ocen. Typowa ocena wynosi około 5 do 6.
- Być może nawet stały klasyfikator, który przewidywałby cały czas średnią ocenę win ze zbioru uczącego, działaby według MAE całkiem dobrze. Sprawdźmy to z ciekawości.
- Średnia ocena dla win ze zbioru uczącego wynosi:

```
> mean(wine_train$quality)
[1] 5.870933
```

KROK 4. OCENA MODELU

- Gdybyśmy dla każdego przykładu ze zbioru testowego przewidzieli tę samą wartość 5.87, to MAE wynosiłby:

```
> MAE(5.87, wine_test$quality)
[1] 0.6722474
```

- Zatem nasz model drzewa regresyjnego jest nieco lepszy (MAE $= 0.59$ vs. MAE $= 0.67$).

- Dla porównania, Cortez i współautorzy (patrz artykuł przytoczony kilka slajdów wcześniej) osiągnęli MAE $= 0.58$ dla sieci neuronowej oraz MAE $= 0.45$ dla SVM (sieci i SVM są modelami typu *black box*, dlatego my próbujemy drzew). Najpewniej istnieje zatem możliwość poprawy naszego modelu.

KROK 5. DOPRACOWANIE MODELU

- W celu poprawy jakości zbudowanego modelu drzewa regresyjnego zbudujemy drzewo modeli, które jest rozszerzeniem metody drzew regresyjnych (poprzez zastąpienie średnich w liściach modelami regresji liniowej).
- Obecnie do budowania drzew modeli wykorzystuje się najczęściej algorytm **Cubist**. Szczegóły implementacyjne tej metody wykraczają poza ramy niniejszego kursu i nie będą tutaj omawiane.
- Algorytm Cubist dostępny jest w R w pakiecie **Cubist** w funkcji `cubist()`.

KROK 5. DOPRACOWANIE MODELU

Składnia metody drzew modeli

z użyciem funkcji `cubist()` z pakietu **Cubist**

Budowa modelu:

```
m <- cubist(train, class)
```

- `train` jest ramką danych lub macierzą zawierającą zbiór uczący
- `class` jest wektorem wartości zmiennej objaśnianej dla przykładów ze zbioru uczącego

Funkcja zwraca model drzewa modeli, który można wykorzystywać do wykonywania predykcji.

KROK 5. DOPRACOWANIE MODELU

Składnia metody drzew modeli

z użyciem funkcji `cubist()` z pakietu **Cubist**

Predykcja:

```
p <- predict(m, test)
```

- `m` to model zbudowany z wykorzystaniem funkcji `cubist()`
- `test` jest ramką danych zawierającą zbiór testowy (z tymi samymi cechami co zbiór uczący)

Funkcja zwraca wektor prognozowanych wartości.

Przykład:

```
wine_model <- cubist(wine_train, wine_quality)
wine_prediction <- predict(wine_model, wine_test)
```

KROK 5. DOPRACOWANIE MODELU

Zadanie 7.

a) Używając algorytmu Cubist, dopasować drzewo modeli do naszych danych.

b) Wyświetlić podstawowe informacje na temat zbudowanego modelu (wywołując zmienną z modelem) oraz strukturę całego drzewa — wszystkie wygenerowane reguły (funkcją `summary()`). Przeanalizować otrzymane wyniki.

KROK 5. DOPRACOWANIE MODELU

Zadanie 8.

a) Za pomocą zbudowanego drzewa modeli wykonać predykcję jakości win ze zbioru testowego.

b) Przeanalizować otrzymane wyniki: wyświetlić podstawowe statystyki wszystkich prognoz (`summary()`), policzyć współczynnik korelacji pomiędzy predykcjami a prawdziwymi wartościami oraz obliczyć średni błąd bezwzględny otrzymanych prognoz.