

Zintegrowany Program Rozwoju
Akademii Górniczo-Hutniczej w Krakowie
Nr umowy: POWR.03.05.00-00-Z307/17

Instrukcja do ćwiczeń laboratoryjnych

Nazwa przedmiotu	Eksploracja danych
Numer ćwiczenia	2
Temat ćwiczenia	Klasyfikacja - drzewa decyzyjne, kNN

Poziom studiów	II stopień
Kierunek	Data Science
Forma studiów	Stacjonarne
Semestr	1

Wojciech Czech

1. Cel ćwiczenia

- Praktyczne zapoznanie się z klasyfikacją za pomocą drzew decyzyjnych oraz klasyfikatorem k -NN
- Przyswojenie pojęć: najbliższy sąsiad, *pruning*, walidacja krzyżowa

2. Wprowadzenie do ćwiczenia

Klasyfikacja w oparciu o drzewa decyzyjne jest jedną z podstawowych metod uczenia maszynowego, przydatną nie tylko w tworzeniu modeli predykcyjnych, ale również w budowaniu baz wiedzy złożonych z łatwych do interpretacji przez człowieka reguł i filtrów na cechach. Drzewa decyzyjne są szczególnie przydatne w przypadku danych nominalnych, danych mieszanych. Znajdują zastosowanie w budowaniu klasyfikatorów złożonych oraz w lasach losowych.

Klasyfikator najbliższego sąsiada jest jedną z najbardziej podstawowych metod klasyfikacji i regresji, pozwalającą na budowę prostych modeli predykcyjnych dla danych osadzonych w przestrzeni wektorowej z różnymi rodzajami metryk.

3. Przykładowe dane

- <http://home.agh.edu.pl/~czech/datasets/ed-titanic-training.csv>
Zbiór danych treningowych na temat ofiar katastrofy Titanica. Zawiera informacje o 890 pasażerach statku wraz z binarną etykietą określającą czy przeżyli katastrofę, czy utonęli. Każdy rekord opisany jest dziesięcioma cechami: *Pclass*, *Sex*, *Age*, *Parch*, *Fare*, *Embarked*, *HasCabin*, *FamilySize*, *IsAlone*, *Title*.
- <http://home.agh.edu.pl/~czech/datasets/ed-titanic-test.csv>
Zbiór danych testowych - pasażerowie Titanica. Zawiera 418 rekordów bez etykiety określającej śmierć w katastrofie. Patrz opis powyżej.
- <http://home.agh.edu.pl/~czech/vis-datasets/misc/nyt-frame.csv>
Zbiór danych o rozmiarze 101×4433 zawierający 101 wektorów cech reprezentujących artykuły New York Times w dwóch kategoriach: muzyka i sztuka. Wektory cech są znormalizowanymi i przeskalowanymi zgodnie z rankingiem IDF wektorami *Bag of Words*. Rozmiar słownika wynosi 4433.
- Breast Cancer Dataset (dostępny w SciKit Learn)

4. Przydatne biblioteki i funkcje

1. Pandas: <https://pandas.pydata.org>
 - `read_csv()`
 - `DataFrame`
2. NumPy:
 - `array`

3. SciKit Learn:
 - PCA
 - DecisionTreeClassifier
 - KFold
 - KNeighborsClassifier
 - RadiusNeighborsClassifier
 - train_test_split()
 - cross_val_score(), confusion_matrix(), f1_score()
4. Seaborn <https://seaborn.pydata.org>

5. Plan ćwiczenia: drzewa decyzyjne

1. Załaduj zbiory danych Titanic (treningowy i testowy), a następnie wyświetl nagłówki z nazwami cech oraz przykładowe rekordy

```
import pandas as pd
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
train.head(3)
```

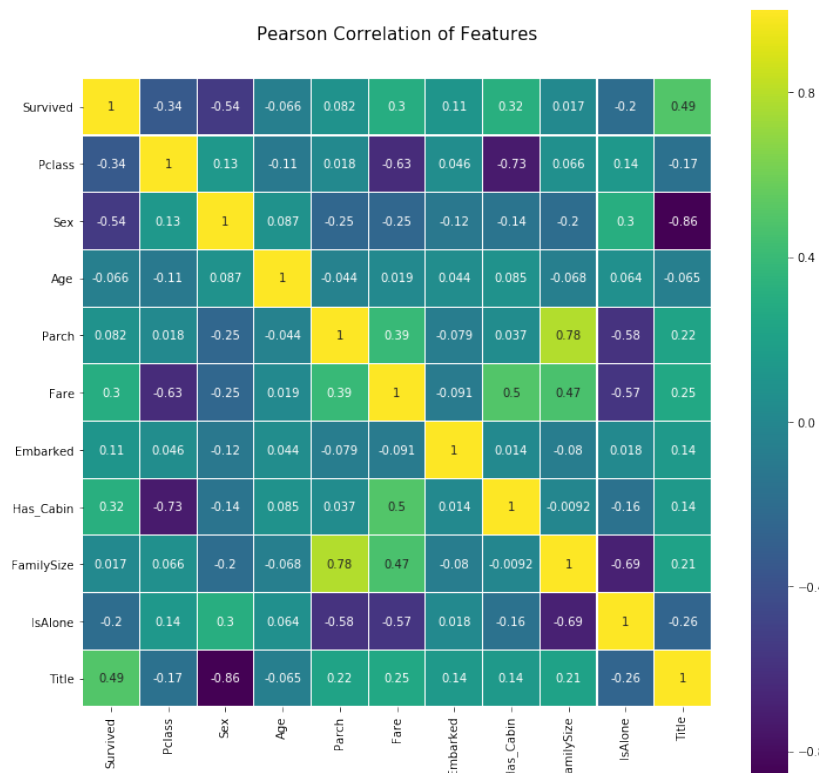
2. Wyznacz korelację Pearsona pomiędzy cechami zbioru treningowego i dokonaj wizualizacji macierzy (patrz Rysunek 1). Które cechy są najbardziej skorelowane z etykietą przeżycia? Które cechy są najbardziej skorelowane ze sobą?

```
import matplotlib.pyplot as plt
import seaborn as sns
colormap = plt.cm.viridis
plt.figure(figsize=(12,12))
plt.title('Pearson Correlation of Features', y=1.05, size=15)
sns.heatmap(train.astype(float).corr(),linewidths=0.1,vmax=1.0,
square=True, cmap=colormap, linecolor='white', annot=True)
```

3. Korzystając ze zbioru treningowego, wyznacz współczynnik przeżywalności dla każdego z pięciu różnych tytułów (cecha *Title*)
4. Korzystając ze zbioru treningowego oraz walidacji krzyżowej (10-fold) wyznacz najlepszą głębokość drzewa decyzyjnego (kryterium podziału *gini*, albo *entropy*)
5. Zbuduj drzewo decyzyjne wykorzystując z wyznaczonej wcześniej maksymalnej głębokości (kryterium podziału *gini*, lub *entropy*)

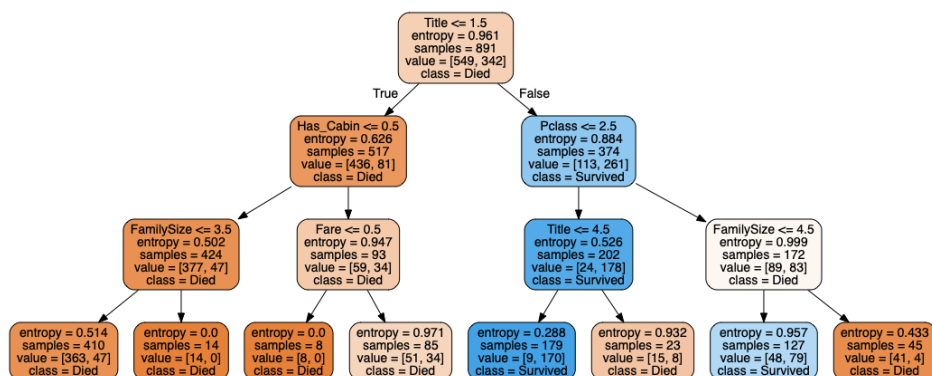
```
from sklearn import tree
decision_tree = tree.DecisionTreeClassifier(max_depth = 3,
criterion='entropy')
decision_tree.fit(x_train, y_train)
```

6. Dokonaj predykcji możliwości przeżycia dla pasażerów ze zbioru testowego i zapisz wyniki w pliku
7. Zwizualizuj drzewo decyzyjne (patrz Rysunek 2)



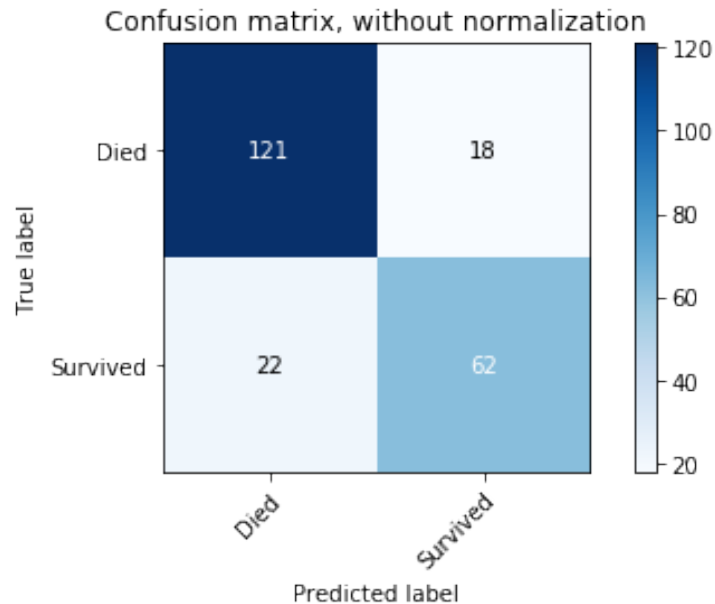
Rysunek 1: Korelacje pomiędzy cechami zbioru treningowego Titanic.

8. Sprawdź jak na dokładność klasyfikacji wpływają następujące parametry drzewa: kryterium podziału (*gini* vs. *entropy*), najmniejsza liczba rekordów w liściu oraz maksymalna głębokość drzewa
9. Podziel zbiór treningowy (dla którego mamy dostępne etykiety) na nowy zbiór treningowy (75%) i nowy zbiór testowy (25%)
10. Wyznacz i dokonaj wizualizacji macierzy rozbieżności *confusion matrix* (patrz Rysunek 3). Wykorzystaj w tym celu pomocniczą funkcję `plot_confusion_matrix()` dostępną w bibliotece SciKit Learn.
11. Wyznacz następujące miary jakości zbudowanego klasyfikatora:
 - *accuracy*



Rysunek 2: Wizualizacja drzewa decyzyjnego o głębokości 3.

- *f1-score*
- *average precision-recall*



Rysunek 3: Przykładowa macierz rozbieżności dla zbioru danych Titanic.

6. Plan ćwiczenia: klasyfikator k -NN

- Przetestuj działanie klasyfikatora najbliższego sąsiada na zbiorach danych: Breast Cancer oraz NYT (wymiarowość zredukowana do 10 za pomocą PCA):
 - Wykorzystując dziesięciokrotną walidację krzyżową zmierz dokładność klasyfikacji dla $k = 1$, $k = 3$, $k = 5$, $k = 7$.
 - Zbadaj wpływ wprowadzenia wag odległości oraz innej miary odległości (Euklidesowa vs. Taxi) na rezultaty klasyfikacji
 - * Zaimplementuj i przetestuj algorytm KD-tree dla n wymiarów

7. Lasy losowe*

- Sprawdź czy zastosowanie wielu nieskorelowanych drzew decyzyjnych (las losowy) może poprawić wyniki klasyfikacji uzyskane w punktach: 4.9, 4.10, 4.11.
- Przetestuj dokładność klasyfikacji uzyskiwaną za pomocą klasyfikatora RandomForest na zbiorze Breast Cancer (rezultaty dziesięciokrotnej walidacji krzyżowej). Jakie wartości parametrów modelu dają najlepsze wyniki? Czy użycie PCA wpływa znacząco na osiągnięte wyniki?

8. Sposób oceny / uzyskania zaliczenia

Na uzyskanie zaliczenia z zajęć laboratoryjnych składa się:

- Wykonanie wszystkich zadań na laboratorium oraz przesłanie kodu za pomocą systemu UPeL

Ocena z zajęć laboratoryjnych (OL , w skali 2 – 5) obliczana jest zgodnie ze wzorem:

$$OL = 0.5 * LA + 0.5 * LW,$$

gdzie:

- LA – ocena aktywności studenta podczas zajęć, wystawiana przez prowadzącego na podstawie zaangażowania studenta w realizację zadań oraz odpowiedzi ustnej na zadane pytania dotyczące realizowanego zadania;
- LW – ocena uzyskana za zadania wykonane na zajęciach (kod źródłowy) wystawiona przez prowadzącego na podstawie poprawności i kompletności zadania przesłanego na platformę UPeL.

9. Literatura

- *Data Mining: The Textbook*, Charu C. Aggarwal, Springer 2015.
- *Data Mining: Concepts and Techniques*, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier 2012, Third Edition.