Sieci neuronowe i sztuczna inteligencja – laboratorium 5

Monika Błyszcz, 236623

W zadaniu zaimplementowano 3 różne testy z kategorii korelacji. Wybrano korelację Pearsona, Spearmana i Kendalla.

- Współczynnik korelacji Pearsona można wykorzystać do podsumowania siły liniowej zależności między dwiema próbkami danych. Oblicza się go jako kowariancję dwóch zmiennych podzieloną przez iloczyn odchylenia standardowego każdej próbki danych. Jest to normalizacja kowariancji między dwiema zmiennymi, aby uzyskać interpretowalny wynik.

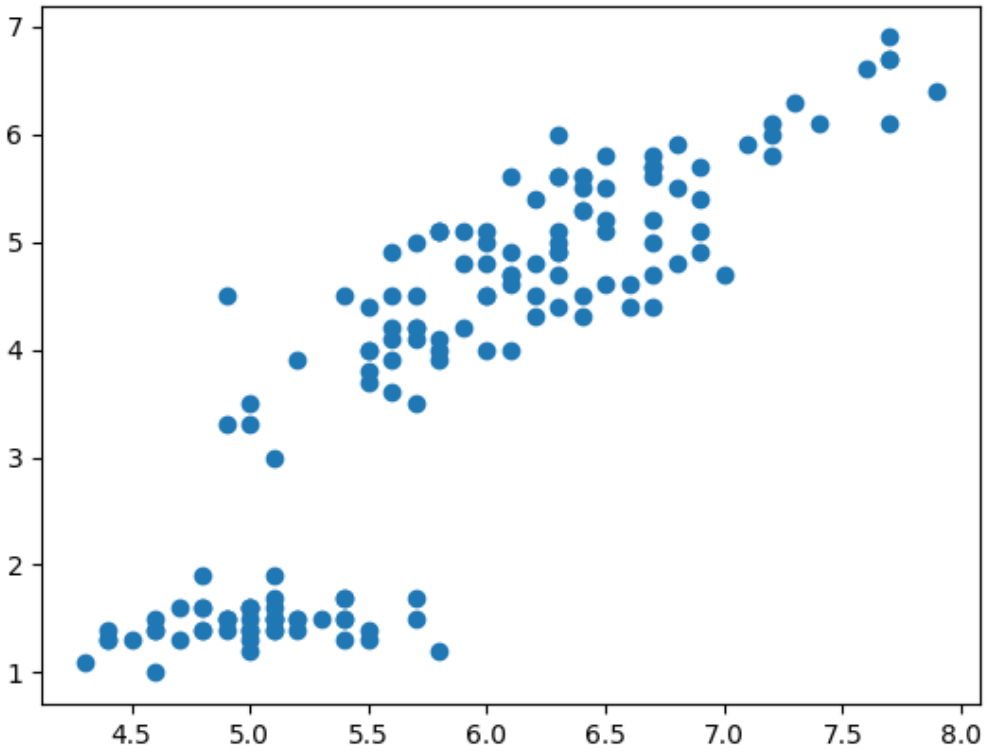
- Współczynnik korelacji Spearmana można wykorzystać do podsumowania siły między dwiema próbkami danych. Zamiast obliczać współczynnik przy użyciu kowariancji i odchyleń standardowych dla samych próbek, statystyki te są obliczane na podstawie względnej rangi wartości na każdej próbce. Jest to powszechne podejście stosowane w statystyce nieparametrycznej, np. metodach statystycznych, w których nie zakładamy rozkładu danych, takich jak gaussowskie. W tym przypadku nie zakłada się liniowej tylko monotoniczną zależność między zmiennymi

- Korelacja rang Kendalla (współczynnik korelacji Kendalla) polega na tym, że oblicza znormalizowany wynik dla liczby pasujących lub zgodnych rankingów między dwiema próbkami. Test przyjmuje dwie próbki danych jako argumenty i zwraca współczynnik korelacji oraz wartość p. Jako test hipotezy statystycznej metoda zakłada (H0), że nie ma związku między dwiema próbkami

Poniżej przedstawiono zaimplementowany kod. Jako bazę danych wykorzystano bazę kwiatów iris_datatset. Szukano korelacji między Sepal Length (długość działki kielicha) a Petal Length (długość płatka).

```python
import pandas
from scipy.stats import pearsonr
from scipy.stats import spearmanr
from scipy.stats import kendalltau
from matplotlib import pyplot

#Load dataset
url="https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv"
names=["sepal-length", "sepal-width", "petal-length", "petal-width", "class"]
dataset = pandas.read_csv(url, names=names)
#data
data1 = dataset.loc[:, 'sepal-length']
data2 = dataset.loc[:, 'petal-length']

#show data
pyplot.scatter(data1, data2)
pyplot.show()

# Pearson's Correlation test
print("Pearson's correlation")
stat, p = pearsonr(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
 print('Probably independent')
else:
 print('Probably dependent')

#Spearman's Rank Correlation Test
print(' ')
print("Spearman's correlation")
stat, p = spearmanr(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
 print('Probably independent')
else:
 print('Probably dependent')

#Kendall's Rank Correlation Test
print(' ')
print("Kendall's correlation")
stat, p = kendalltau(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
 print('Probably independent')
else:
 print('Probably dependent')
```

Najpierw wykreślono zależność na wykresie: widzimy, że prawie każda zmienna ma równomierny rozkład, a dodatni związek między zmiennymi jest widoczny przez ukośne grupowanie punktów od lewego dolnego do prawego górnego rogu wykresu. Zależność nie jest do końca liniowa, bo w dolnym lewym rogu występują odseparowane dane. Odrzucenie tych danych sprawi, że zależność będzie liniowa.



Otrzymano następujące wyniki:

```
C:\Users\Mo\AppData\Local\Microsoft\WindowsApps\python3.10.exe C:\Users\Mo\
Pearson's correlation
stat=0.872, p=0.000
Probably dependent

Spearman's correlation
stat=0.881, p=0.000
Probably dependent

Kendall's correlation
stat=0.718, p=0.000
Probably dependent

Process finished with exit code 0
```

Analizując powyższe wyniki widzimy, że występuje dość duża, pozytywna korelacja pomiędzy długością kielicha a długością płatka. Dla poszczególnych testów statystycznych mamy:

- **Korelacja Pearsona:** obie zmienne są dodatnio skorelowane i że korelacja wynosi 0,872. Sugeruje to wysoki poziom korelacji pomiędzy wartościami
- **Korelacja Spearmana:** test statystyczny wykazuje silną dodatnią korelację z wartością 0,881. Wartość p równa zeru, co oznacza, że prawdopodobieństwo zaobserwowania danych, uwzględniając, że próbki są nieskorelowane, jest bardzo mało prawdopodobne. Pozwala to na odrzucenie hipotezy zerowej, że próbki są nieskorelowane.
- **Korelacja rang Kendalla:** wyniosła 0,718, co oznacza, że zależność jest wysoce skorelowana. Wartość p jest bliska zeru (i wydrukowana jako zero), podobnie jak w korelacji Spearmana, co oznacza, że możemy odrzucić hipotezę zerową, że próbki są nieskorelowane.

Najwyższy współczynnik korelacji pokazała korelacja Spearmana, co jest zgodne z rzeczywistością, bo dane nie są w pełni liniowe.