Laboratorium 2 — Metody preprocesingu

Normalizacja jest jedną z procedur wstępnej obróbki danych w celu umożliwienia ich wzajemnego porównania i dalszej analizy.

Zadanie 1

Utwórz poniższą macierz:

$$a = \begin{bmatrix} 1 & 2 \\ 5 & 6 \\ -2 & 2 \\ 3 & 7 \\ 2 & 2 \\ 3 & -6 \\ -1 & -2 \\ -2 & 8 \\ 4 & 1 \\ -3 & 3 \\ 0 & 4 \\ 1 & -9 \end{bmatrix}$$

Zadanie 2 - normalizacja zachowująca zero

Utwórz nową macierz zawierającą znormalizowane dane (pamiętaj aby każdą kolumnę/atrybut potraktować niezależnie) za pomocą następującego wzoru

$$x_{norm} = \frac{x}{max|x|}.$$

Wyświetl minimalną, maksymalną oraz średnią wartość i odchylenie standardowe dla każdego atrybutu.

Zadanie 3 - skalowanie

Utwórz nową macierz zawierającą przeskalowane dane (pamiętaj aby każdą kolumnę/atrybut potraktować niezależnie) za pomocą następującego wzoru

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)}.$$

Wyświetl minimalną, maksymalną oraz średnią wartość i odchylenie standardowe dla każdego atrybutu.

Zadanie 4 - standaryzacja

Utwórz nową macierz zawierającą przeskalowane dane (pamiętaj aby każdą kolumnę/atrybut potraktować niezależnie) za pomocą następującego wzoru

$$x_{norm} = \frac{x - mean(x)}{std(x)}.$$

Wyświetl minimalną, maksymalną oraz średnią wartość i odchylenie standardowe dla każdego atrybutu. Na co wpływa standaryzacja?

Binaryzacja jest kolejną z procedur wstępnej obróbki danych, pozwala ona przekształcić atrybut nominalny na jedene lub więcej atrybutów binarnych.

Zadanie 5

Utwórz poniższą macierz:

$$a = \begin{bmatrix} red & yes \\ red & yes \\ green & no \\ blue & no \\ blue & no \\ green & yes \\ red & no \\ red & no \\ yellow & yes \end{bmatrix}$$

Zadanie 6

Dokonaj binaryzacji powyższej macierzy. Jaka liczba atrybutów została utworzona w wyniku binaryzacji atrybutu 1? Jaka jest optymalna liczba kolumn powstała w wyniku binaryzacji atrybutu 2?

Selekcja atrybutów jest kolejną techniką wstępnej obróbki danych, jej zadaniem jest pozbycie się atrybutów nieistotnych z punktu widzenia danego zadania w celu zmniejszenia wymiarowości danych, lub umożliwienie wizualizacji posiadanych danych. Gotowe algorytmy selekcji cech możemy znaleźć między innymi w pakiecie scikit-learn.

Zadanie 7

Wczytaj zbiór iris, a następnie korzystając z klasy VarianceThreshold z pakietu scikit-learn usuń atrybuty o niskiej wariancji, za próg przycinania przyjmij 0.2 (pamiętaj, aby pominąć kolumnę z klasą). Ile kolumn zostało usuniętych?

Zadanie 8

Wczytaj zbiór iris, a następnie korzystając z klasy Mutual_info_classif z pakietu scikit-learn wyznacz wartość informacji wzajemnej dla poszczególnych atrybutów. Wybierz 3 najistotniejsze atrybuty.

Zadanie 9

Wczytaj zbiór iris, a następnie korzystając z klasy PCA z pakietu scikit-learn zmniejsz wymiarowość do 2, a następnie utwórz wykres stosując pakiet matplotlib.