Laboratorium maszynowej analizy danych

Laboratorium 1

Przygotowanie danych do analizy z wykorzystaniem uczenia maszynowego.

Wprowadzenie

Model – część systemu uczenia maszynowego odpowiedzialna za uczenie się i uzyskiwanie przewidywań.

Zbiór/zestaw treningowy/uczący (ang. *training set*) – przykładowe dane używane do trenowania systemu (model trenowany jest za pomocą zbioru uczącego).

Zbiór/zestaw testowy (ang. *test set*) – zbiór za pomocą którego sprawdzany jest model uczenia maszynowego.

Ogólny schemat postępowania przy projekcie uczenia maszynowego:

- zdefiniowanie celu/problemu do rozwiązania
- **pozyskanie i przygotowanie danych**
- wybór modelu
- wytrenowanie modelu na danych uczących
- dostrojenie modelu
- wykorzystanie wytrenowanego modelu do prognozowania wyników dla nowych przypadków
- wdrażanie, monitorowanie, utrzymanie i konserwacja systemu

Główne problemy uczenia maszynowego:

- **niedobór danych uczących**
- **niereprezentatywne dane uczące**
- **dane kiepskiej jakości**
- **nieistotne cechy**
- przetrenowanie danych uczących (ang. *overfitting*)
- niedotrenowanie danych uczących (ang. *underfitting*)

Przygotowanie danych

1. Pobranie danych.
2. Analiza struktury danych i wizualizacja w celu rozpoznania wzorców i dodatkowych informacji.
3. **Czyszczenie, formatowanie, usuwanie lub oszacowanie brakujących wartości w danych:**
    a. Brakujące wartości:
        i. Usunięcie przykładów zawierających brakujące dane,
        ii. Usunięcie całego atrybutu,
        iii. Imputacja (uzupełnienie brakujących danych określoną wartością).
    b. Obsługa tekstu i atrybutów kategorialnych:
        i. Przekształcenie kategorii z tekstu na wartości numeryczne
        ii. Kodowanie „gorącojedynkowe”/ dummying
        iii. Embedding
    c. Dyskretyzacja cech
    d. Niezrównoważone klasy
    e. Punkty odstające

**drzewa decyzyjne, lasy losowe, boosting gradientowy są bardziej odporne na punkty odstające*

4. **Skalowanie danych (ang. *feature scaling*):**
    a. Standaryzacja:
        i. polega na przekształceniu danych pierwotnych, aby ich rozkład miał średnią wartość równą 0 i odchylenie standardowe równe 1,
        ii. nie jest ograniczona do określonego zakresu,
        iii. rozkład cech jest normalny lub gaussowski,
        iv. mniej wrażliwa na elementy odstające,
        v. często stosowane w modelach liniowych (regresja liniowa, logistyczna), PCA, itd.
    b. Normalizacja:
        i. stosowana, gdy dane mają różne wymiary,
        ii. wszystkie cechy są tej samej skali, **najczęściej*** od 0 do 1 lub -1 do 1,
        iii. stosowana, gdy rozkład cech jest nieznany lub nie jest gaussowski,
        iv. wrażliwa na elementy odstające,
        v. często stosowana w algorytmach wykorzystujących odległość (np. knn, k-średnich, SVM), sieciach neuronowych.

        *** w przypadku analizy różnego typu danych pomiarowych, np. widm, dane można normalizować do maksimum, do konkretnej wartości, do pola powierzchni pod krzywą, itp.**

    * rozkład cechy gruboogonowy:
    zastąpienie wartości cechy np. logarytmami,
    kubełkowanie (ang. *bucketizing*) cechy
    * rozkład cechy wielomodalny:
    kubełkowanie,
    dodawanie cechy dla każdego z modów

    c. Niestandardowe transformatory

5. **Grupowanie lub selekcja.**
6. Analiza korelacji.
7. *Tworzenie kombinacji atrybutów.*

!!! Dane wykorzystane do trenowania modelu uczenia maszynowego:

- reprezentatywne,
- usystematyzowane,
- odpowiednia ilość,
- ujednolicony zapis.

Wdrażanie, monitorowanie, utrzymanie i konserwacja systemu

Gdy model uczenia maszynowego jest wszechstronnie przetestowany i zoptymalizowany, dobrze generalizuje wyniki, to należy przygotować go do warunków produkcyjnych (wdrażanie, monitorowanie, utrzymanie i konserwacja systemu).

Cel ćwiczenia

Celem ćwiczenia jest dobór odpowiedniego sposobu skalowania i jego przeprowadzenie dla różnych zestawów danych.

Informacje dodatkowe

Standaryzacja

$$X'_{standarized} = \frac{X - \mu}{\sigma}, \qquad \mu = 0, \qquad \sigma = 1$$

Normalizacja

$$X'_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} (New\ max - New\ min) + New\ min$$

Zadania

1. Proszę o pobranie danych ze źródła: https://archive.ics.uci.edu/dataset/53/iris

   Zbiór danych dotyczy 3 gatunków irysów. Każdy irys jest opisany za pomocą 4 cech (długość i szerokość kielicha, długość i szerokość płatka) + informacja o gatunku:

   1.1. Należy wczytać dane.
   1.2. Sprawdzić podstawowe statystyki.
   1.3. Sprawdzić kompletność danych.
   1.4. Przeprowadzić normalizację danych w zakresie [0,1] na kolumnach (z wyłączeniem kolumny species).
   1.5. Przeprowadzić normalizację danych w zakresie [-1,1] na kolumnach (z wyłączeniem kolumny species)
   1.6. Przeprowadzić standaryzację danych na kolumnach (z wyłączeniem kolumny species).

1.7. Narysować wykres zbiorczy, składający się z 4 wykresów, zależności długości płatka [cm] od szerokości płatka [cm] różnicowany na podstawie gatunku dla danych pierwotnych, znormalizowanych w zakresie [0,1], znormalizowanych w zakresie [-1,1] oraz standaryzowanych.
1.8. Narysować wykres zbiorczy, składający się z 4 wykresów, zależności długości kielicha [cm] od szerokości kielicha [cm] różnicowany na podstawie gatunku dla danych pierwotnych, znormalizowanych w zakresie [0,1], znormalizowanych w zakresie [-1,1] oraz standaryzowanych.

2. Proszę o wczytanie pliku *Zad2_L1.csv*

   Zbiór danych dotyczy serii zarejestrowanych pomiarów w postaci widm Ramana w czasie. Widmo Ramana to wykres intensywności rozproszonego promieniowania Ramana w funkcji różnicy częstotliwości w stosunku do promieniowania padającego. W pierwszym wierszu znajdują się jednostki (Wavenumber [cm^-1], Intensity [a.u.]), w drugim oznaczenia kolejnych pomiarów, t_i. Pierwsza kolumna odnosi się do zakresu pomiarowego, w którym rejestrowane były widma (oś horyzontalna). Pozostałe kolumny odnoszą się do poszczególnych pomiarów. Należy wykonać następujące operacje:

   2.1. Przeprowadzić normalizację danych: każde widmo należy znormalizować do amplitudy pasma przy 985 cm^{-1} (z wyłączeniem kolumny Wavenumber [cm^-1]).
   2.2. Narysować wykres zbiorczy składający się z 2 wykresów: widma dla danych surowych oraz widma znormalizowane.
   2.3. Jaka jest przyczyna zastosowania powyższego sposobu normalizacji?

3. Proszę o wczytanie pliku *Zad3_L1.csv*

   Zbiór danych dotyczy serii zarejestrowanych pomiarów w postaci widm FTIR w czasie. Widmo w podczerwieni obrazuje intensywność widma w podczerwieni. Pierwsza kolumna odnosi się do zakresu pomiarowego (oś pozioma), w którym rejestrowane były widma (Wavenumber [cm^-1]). Pozostałe kolumny (Absorbance [a.u.])) odnoszą się do widm rejestrowanych po upływie określonego czasu (oś wertykalna). Należy wykonać następujące operacje:

   3.1. Przeprowadzić normalizację danych: każde widm należy znormalizować do pola powierzchni pod wykresem (z wyłączeniem kolumny Wavenumber [cm^-1]).
   3.2. Narysować wykres zbiorczy składający się z 2 wykresów: widma dla danych surowych oraz widma znormalizowane.
   3.3. Jaka jest przyczyna zastosowania powyższego sposobu normalizacji?

Literatura

1. A. Géron, *Uczenie maszynowe z użyciem Scikit-Learn, Keras, i TensorFlow,* Wydanie III, Helion, 2023.
2. S.J. Russell, P. Norvig, *Sztuczna inteligencja. Nowe spojrzenie,* Wydanie IV, Helion, 2023.
3. A. Król-Nowak, K. Kotarba, *Podstawy uczenia maszynowego,* Wydawnictwa AGH, Kraków, 2022.
4. K. Gallatin, K. Albon, *Uczenie maszynowe w Pythonie. Receptury. Od przygotowania danych do deep learningu*, Wydanie II, Helion, 2023.
5. J. VanderPlas, *Python Data Science. Niezbędne narzędzia do pracy z danymi,* Wydanie II, Helion, 2023.