POLITECHNIKA GDAŃSKA

AI TECH

Wprowadzenie do Sztucznej Inteligencji

PROJEKT - WPROWADZENIE

prof. dr hab. inż. Bożena Kostek (p. 731)
LAF/KSM WETI

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Wprowadzenie do sztucznej inteligencji – projekt (instrukcja)

Bożena Kostek

wprowadzenie_SI@multimed.org

Cel projektu

- Celem projektu jest przygotowanie systemu wnioskowania, wykorzystującego wybrane algorytmy sztucznej inteligencji;
- Nabycie umiejętności pogłębionej analizy danych (w tym medycznych).

Organizacja zajęć (zajęcia nr 2/3, 14.04.22 r. /21.04.22 r.)

- Projekt – wprowadzenie 14.04.22 r.
- **Zajęcia regularne** (II połowa semestru):
  - **czwartki, godz.** 13:15-14:45
  - **czwartki**, godz. 15:00-16:45

Zajęcia projektowe: 21.04.22 r. g. 13:15

Wprowadzenie do systemu R – prof. Piotr Szczuko (1 h)

Wprowadzenie do systemu WEKA – dr Michał Lech (2 h)

Projekt – wytyczne ogólne

Opis zagadnienia, środowiska: R, WEKA, (C++, Java, Python, Matlab, biblioteki)
Wprowadzenie do projektu. Cele projektu (BK)
Wprowadzenie do systemu R (PS) Wprowadzenie do systemu WEKA (ML)
Wytyczne: Wybór i zapoznanie się z wybraną bazą danych medycznych i in.; Zapoznanie się z jednostką chorobową i rodzajem danych, które opisują daną chorobę
Krytyczna analiza danych zawartych w bazie; Wstępne przygotowanie danych (np. parametryzacja, normalizacja, dyskretyzacja danych, redukcja danych wybranymi metodami, np. metoda PCA (Principle Component Analysis, metoda głównych składowych);

Projekt - wytyczne ogólne

Opis zagadnienia
Podział danych na zbiory: treningowe, walidacyjne, testowe; Wybór klasyfikatora/ów (systemu, środowiska) – z uzasadnieniem (śr., algorytm)
Klasyfikacja zbioru testowego; Przedstawienie uzyskanych wyników; Analiza uzyskanych wyników (skuteczność – macierz pomyłek, analiza wyników walidacji krzyżowej, inne wskaźniki klasyfikacji i ich interpretację); Pokazanie dalszych kierunków rozwoju przygotowanego systemu wnioskowania (klasyfikacji danych medycznych) – krytyczna analiza uzyskanych wyników klasyfikacji;
Przygotowanie opracowania pisemnego (prezentacje .ppt)

Zadania projektowe (zajęcia nr 2, 14.04.22 r., 2h)

Projekt obejmuje:

- Wybór i zapoznanie się z wybraną bazą danych (w tym medycznych);
- Zapoznanie się z danymi; jednostką chorobową i rodzajem danych, które opisują daną chorobę;
- Zapoznanie się z publikacjami, które odnoszą się do danej bazy (w tym do *state-of-the-art*);
- Krytyczną analizę danych zawartych w bazie;

Zadania projektowe

Projekt obejmuje:

- Wstępne przygotowanie danych (np. parametryzacja, normalizacja, dyskretyzacja danych, redukcja danych wybranymi metodami, np. metoda PCA (*Principle Component Analysis*, metoda głównych składowych) – porównanie z zast. i bez zast.;
- Podział danych na zbiory: treningowe, walidacyjne, testowe;
- Wybór klasyfikatora/ów (systemu) – z uzasadnieniem;

Zadania projektowe

Projekt obejmuje:

- Klasyfikację zbioru testowego;
- Przedstawienie uzyskanych wyników;
- Analizę uzyskanych wyników (skuteczność – macierz pomyłek, analiza wyników walidacji krzyżowej, **inne wskaźniki klasyfikacji i ich interpretację**);

Zadania projektowe

Projekt obejmuje:

- Pokazanie dalszych kierunków rozwoju przygotowanego systemu wnioskowania (klasyfikacji danych medycznych) - – krytyczna analiza uzyskanych wyników klasyfikacji;
- W sprawozdaniu pisemnym należy podać odwołania do bibliografii
- Prezentację wyników oraz przygotowanie sprawozdania z przebiegu projektu.

Analiza danych – wskaźniki jakości klasyfikacji

- TP – *True Positive* – liczba obserwacji poprawnie zaklasyfikowanych do klasy pozytywnej
  TN – *True Negative* – liczba obserwacji poprawnie zaklasyfikowanych do klasy negatywnej

Analiza danych – wskaźniki jakości klasyfikacji

- FP – *False Positive* – liczba obserwacji zaklasyfikowanych do klasy pozytywnej podczas, gdy w rzeczywistości pochodzą z klasy negatywnej
  FN – *False Negative* – liczba obserwacji zaklasyfikowanych do klasy negatywnej podczas, gdy w rzeczywistości pochodzą z klasy pozytywnej

Analiza danych – wskaźniki jakości klasyfikacji

- TPR (True Positive Rate) – określa zdolność klasyfikatora do wykrywania klasy pozytywnej (stanu patologicznego)
- TPR = TP / (TP + FN)
- TNR (True Negative Rate) – określa zdolność klasyfikatora do wykrywania klasy negatywnej (stanu normalnego)
- TNR = TN / (TN + FP)

Analiza danych – wskaźniki jakości klasyfikacji

- FPR (False Positive Rate) – określa, jak często klasyfikator popełnia błąd, klasyfikując stan normalny jako patologiczny
- FPR = FP / (FP + TN)
- FNR (False Negative Rate) – określa, jak często klasyfikator popełnia błąd, klasyfikując stan patologiczny jako normalny

Analiza danych – wskaźniki jakości klasyfikacji

- SE (czułość (*sensitivity*), ***Recall*** (*True Positive Value*)) – określa zdolność klasyfikatora do wykrywania klasy pozytywnej (stanu patologicznego)

  SE = TP / (TP + FN)

- SP (specyficzność (*specificity*), *True Negative Rate*) – określa zdolność klasyfikatora do wykrywania klasy negatywnej (stanu normalnego)

  SP = TN / (TN + FP)

Analiza danych – wskaźniki jakości klasyfikacji

- ***Precision*** (precyzja, PPV – *Postive Predictive Value*) określa, jaka część wyników wskazanych przez klasyfikator jako dodatnie jest faktycznie dodatnia

  PPV = TP/TP+FP

Analiza danych – wskaźniki jakości klasyfikacji

- ACC (*Total Accuracy*) – dokładność, całkowita sprawność klasyfikatora, określa prawdopodobieństwo poprawnej klasyfikacji, czyli stosunek poprawnych klasyfikacji do wszystkich klasyfikacji

  ACC = (TP + TN) / (TP + TN + FP + FN)

- Przy czym, istnieją takie zależności:

  SE = TPR 1 – SE = FNR

  SP = TNR 1-SP = FPR

Analiza danych – wskaźniki jakości klasyfikacji

- Miara *F – measure* jest średnią ważoną precyzji (*Precision*) i rozrzutu (*Recall*). Stanowi kombinację obu miar. We wzorze została przedstawiona formuła opisująca miarę F.

$$F1\ score = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

Analiza danych – wskaźniki jakości klasyfikacji

- ROC (ang. *Receiver Operating Characteristic*) - ocena jakości klasyfikacji
- Krzywa ROC ilustruje związek pomiędzy czułością a specyficznością dla danego modelu
- Bardzo popularnym podejściem jest obliczenie pola pod krzywą ROC – AUC (*Area Under Curve*) i traktowanie go jako miarę dobroci i trafności danego modelu

Zadania projektowe – redukcja danych

- Obsługa wartości pustych (ang. *missing values, null values*) oraz oddalonych (ang. *outliers*) jest niezmiernie ważna w procesie redukcji wymiarowości.
- Wiele z metod redukcji wymiarowości albo w ogóle nie zadziała w obecności pustych wartości, albo otrzymane wyniki będą zniekształcone poprzez „oddalenie” danych.
- Nieuwzględnienie z kolei w analizie eksploracyjnej części danych (odrzucenie wierszy ze „niepoprawnymi” danymi) jest ze statystycznego punktu widzenia dyskusyjne.

Bazy danych (w tym medycznych)

- h**https://archive-beta.ics.uci.edu/**
- http://archive.ics.uci.edu/ml/datasets/Echocardiogram
- http://archive.ics.uci.edu/ml/datasets/Heart+Disease
- http://archive.ics.uci.edu/ml/datasets/Hepatitis
- http://archive.ics.uci.edu/ml/datasets/Liver+Disorders
- http://archive.ics.uci.edu/ml/datasets/Lung+Cancer
- http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
- https://hbiostat.org/data/

Grafik zawarty na stronie: https://docs.google.com/document/d/1lrLWYBz0aZa_fTDb2j01MpxeYjC0KfsYPEquXv4-eQg/edit

Bazy danych (w tym medycznych)

- h**https://archive-beta.ics.uci.edu/**
- http://archive.ics.uci.edu/ml/datasets/Echocardiogram
- http://archive.ics.uci.edu/ml/datasets/Heart+Disease
- http://archive.ics.uci.edu/ml/datasets/Hepatitis
- http://archive.ics.uci.edu/ml/datasets/Liver+Disorders
- http://archive.ics.uci.edu/ml/datasets/Lung+Cancer
- http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
- https://hbiostat.org/data/

Grafik zawarty na stronie: https://docs.google.com/document/d/1lrLWYBz0aZa_fTDb2j01MpxeYjC0KfsYPEquXv4-eQg/edit

Bazy danych (w tym medycznych)

- https://dl.acm.org/doi/10.1145/1401890.1402012 Image, clinical, and genetic data for Alzheimer's disease (free/on request)
- http://www.cdc.gov/brfss/ Behavioral Risk Factor Surveillance System (BRFSS)
-
- https://www.nlm.nih.gov/hsrph.html A large collection of data sets on various aspects of health care hosted by the NIH
-

Grafik zawarty na stronie: https://docs.google.com/document/d/1lrLWYBz0aZa_fTDb2j01MpxeYjC0KfsYPEquXv4-eQg/edit

Prezentacje projektowe

I prezentacja powinna zawierać:

- Wybór i zapoznanie się z wybraną bazą danych (medycznych);
- Odniesienie się do źródeł opisujących bazę oraz do *state-of-the-art*
- Zapoznanie się z danymi; jednostką chorobową i rodzajem danych, które opisują daną chorobę;
- Wstępne przygotowanie danych (np. parametryzacja, normalizacja, dyskretyzacja danych, redukcja danych wybranymi metodami, np. metoda PCA (*Principle Component Analysis*, metoda głównych składowych);

Prezentacje projektowe

I prezentacja powinna zawierać:

- Krytyczną analizę danych zawartych w bazie;
- Wybór klasyfikatora/ów/systemu – uzasadnienie
- Odniesienia do źródeł literatury

Prezentacje projektowe

II prezentacja powinna zawierać:

- Podział danych na zbiory: treningowe, walidacyjne, testowe;
- Klasyfikację zbioru testowego;
- Przedstawienie uzyskanych wyników;
- Analizę uzyskanych wyników (skuteczność – macierz pomyłek, analiza wyników walidacji krzyżowej, inne wskaźniki klasyfikacji);

Prezentacje projektowe

II prezentacja powinna zawierać:

- Pokazanie dalszych (możliwych) kierunków rozwoju przygotowanego systemu wnioskowania (klasyfikacji danych medycznych) – krytyczna analiza uzyskanych wyników klasyfikacji (dane, wybór klasyfikatora, wyniki).

POLITECHNIKA GDAŃSKA

AI TECH

Dziękuję

Bożena Kostek

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.