Wprowadzenie do sztucznej inteligencji – projekt (instrukcja)

Bożena Kostek

Email: wprowadzenie_SI@multimed.org

Cel projektu

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

Organizacja zajęć (16.04.25)

- Projekt – wprowadzenie **16.04.25**r.
- **Zajęcia** (II połowa semestru):

- Środy godz. 16:15-18/czwartki, godz. 12:15-14:00/14:15-16 – (sala 736), ew. piątek 14:15-16

Zamiana sem. ze środy 29.04 -> czw. 23.04

Zajęcia projektowe dla całej grupy

Kolejne zajęcia będą zawarte wg grafiku na linku, w którym należy się dopisać do danego terminu i grupy:

https://docs.google.com/document/d/1lrLWYBz0aZa_fTDb2j01MpxeYjC0KfsYPEquXv4-eQg/edit?tab=t.0

Organizacja zajęć (16.04.25)

GR.	29.04 środa (zamiana z sem.) 29.04->23.04	13.05 środa	14.05 czwartek	20.05 środa	21.05 czwartek	28.05 czwartek (PSz)
1	x			x		
2	x			x		
3		x		x		
4		x			x	
5		x			x	
6			x			x
7			x			x

Alternatywnie:

29.04	**8.05**	**13.05**	**14.05**	**20.05**	**21.05**	~~**28.05**~~
śr.	**piątek**	**środa**	**czw.**	**śr.**	**czw.**	~~**czw.(PS)**~~
16-18	14-16	16-18	12-14	16-18	12-14	~~12-14~~

Projekt – wytyczne ogólne

Opis zagadnienia, możliwe środowiska/języki programowania: R (WEKA, C++, Java, Python, Matlab, biblioteki)

Wprowadzenie do projektu. Cele projektu (BK)

Wytyczne: Wybór i zapoznanie się z wybraną bazą danych medycznych; 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), wizualizacja danych;

Projekt - **wytyczne ogólne**

Opis zagadnienia

Podział danych na zbiory: treningowe, walidacyjne, testowe; Wybór klasyfikatorów (systemu, środowiska) – z uzasadnieniem (środowisko, algorytmy)

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

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., wizualizacja danych;
- Podział danych na zbiory: treningowe, walidacyjne, testowe;
- Wybór klasyfikatorów (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:

  $$SE = TPR \quad 1 – SE = FNR$$

  $$SP = TNR \quad 1\text{-}SP = FPR$$

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 (przykłady, w tym medyczne)

- 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 wystąpień zawarty na linku (poniżej inf. dotyczącej seminarium):
https://docs.google.com/document/d/1lrLWYBz0aZa_fTDb2j01MpxeYjC0KfsYPEquXv4-eQg/edit?tab=t.0

Bazy danych (przykłady, w tym medyczne)

- 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

Prezentacje projektowe

I prezentacja powinna zawierać:

- Wybór i zapoznanie się z wybraną bazą danych (medycznych);
- Zapoznanie się z danymi; jednostką chorobową i rodzajem danych, które opisują daną chorobę;
- Odniesienie się do źródeł opisujących bazę oraz do *state-of-the-art*
- 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).

Dziękuję

Bożena Kostek