UNIWERSYTET PRZYRODNICZY WE WROCŁAWIU

Data Mining
Wykład 3

Klasyfikacja danych
Klasyfikacja poprzez indukcje drzew decyzyjnych

Plan wykładu

- Sformułowanie problemu
- Kryteria oceny metod klasyfikacji
- Metody klasyfikacji
- Klasyfikacja poprzez indukcje drzew decyzyjnych

Co to jest klasyfikacja

- Polega ona na **znajdowaniu odwzorowania danych w zbiór predefiniowanych klas.**
- Budowany jest model (np. drzewo decyzyjne, reguły logiczne), który służy **do klasyfikowania nowych obiektów** lub głębszego **zrozumienia istniejącego podziału** obiektów na predefiniowane klasy.
- Klasyfikacja jest **metoda eksploracji danych z nadzorem** (z nauczycielem).
- Proces klasyfikacji składa się z kilku etapów:
  1. **Budowa modelu,**
  2. **Walidacja modelu,**
  3. **Testowanie modelu,**
  4. **Predykcji nieznanych wartości.**

Klasyfikacja - Założenia

- Dane wejściowe:

  Treningowy zbiór krotek (przykładów, obserwacji, próbek), będących lista wartości atrybutów opisowych (tzw. deskryptorów) i wybranego atrybutu decyzyjnego (*ang. class label attribute*)

- Dane wyjściowe:

  Model (klasyfikator), przydziela każdej krotce wartość atrybutu decyzyjnego w oparciu o wartości pozostałych atrybutów (deskryptorów)

Klasyfikator

- Wartości atrybutu decyzyjnego dzielą zbiór krotek na predefiniowane klasy, składające się z krotek o tej samej wartości atrybutu decyzyjnego.

Klasyfikator - Służy do predykcji wartości atrybutu decyzyjnego (klasy) krotek, dla których wartość atrybutu decyzyjnego, tj. przydział do klasy, nie jest znany

Proces klasyfikacji

- Klasyfikacja danych jest 4-etapowym procesem:
  - **Etap 1:**

    Budowa modelu (klasyfikatora) opisującego predefiniowany zbiór klas danych lub zbiór pojęć
  - **Etap 2:**

    Walidacja modelu (klasyfikatora) na pewnej części danych treningowych
  - **Etap 3:**

    Testowanie modelu (klasyfikatora) na nowych danych testowych
  - **Etap 4:**

    Zastosowanie opracowanego modelu do klasyfikacji nowych danych

Proces uczenia i testowania

- Zbiór dostępnych krotek (przykładów, obserwacji, próbek) dzielimy na dwa zbiory: **zbiór treningowy** i **zbiór testowy**
- Model klasyfikacyjny (klasyfikator) jest budowany trzyetapowo:

Uczenie (trening) – klasyfikator jest budowany w oparciu o zbiór treningowy danych

Walidacja modelu (klasyfikatora) na pewnej części danych treningowych

Testowanie – dokładność (jakość) klasyfikatora jest weryfikowana w oparciu o zbiór testowy danych

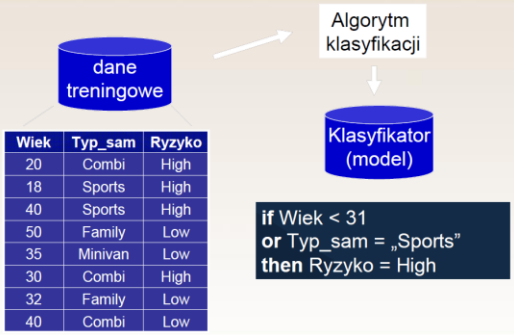
Proces uczenia i testowania

- Wynik klasyfikacji:
  - Reguły klasyfikacyjne postaci if – then
  - Formuły logiczne
  - Drzewa decyzyjne
- Dokładność modelu:

Dla przykładów testowych, dla których znane są wartości atrybutu decyzyjnego, wartości te są porównywane z wartościami atrybutu decyzyjnego generowanymi dla tych przykładów przez klasyfikator

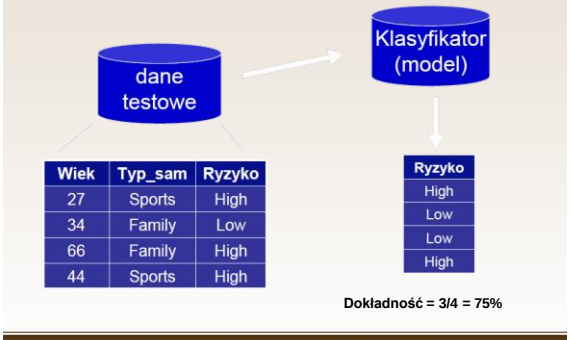
Współczynnik dokładności (*ang. accuracy rate*) = procent przykładów testowych poprawnie zaklasyfikowanych przez model

Uczenie



Wiek	Typ_sam	Ryzyko
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low

Testowanie

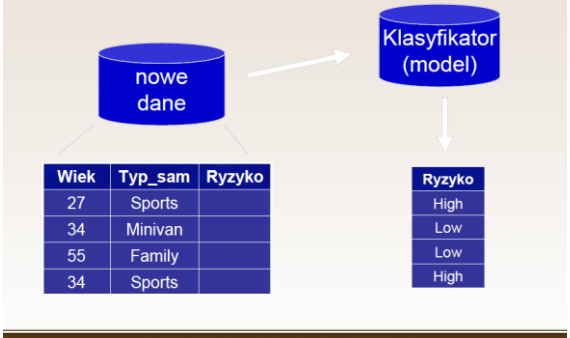


Wiek	Typ_sam	Ryzyko
27	Sports	High
34	Family	Low
66	Family	High
44	Sports	High

Ryzyko
High
Low
Low
High

Dokładność = 3/4 = 75%

Klasyfikacja (predykcja)



Wiek	Typ_sam	Ryzyko
27	Sports	
34	Minivan	
55	Family	
34	Sports	

Ryzyko
High
Low
Low
High

Klasyfikacja a predykcja

- Dwie metody, które są stosowane do analizy danych i ekstrakcji modeli opisujących klasy danych lub do predykcji trendów:
  - **klasyfikacja:**

    predykcja wartości atrybutu kategorycznego (predykcja klasy)

  - **predykcja:**

    modelowanie funkcji ciągłych

Kryteria oceny metod klasyfikacji

- Trafność klasyfikacji (ang. Classification / Predictive accuracy)

  zdolność modelu do poprawnej predykcji wartości atrybutu decyzyjnego (klasy) nowego przykładu

- Szybkość i skalowalność (ang. Speed):
  - czas uczenia się,
  - szybkość samego klasyfikowania

  koszt obliczeniowy związany z wygenerowaniem i zastosowaniem klasyfikatora

- Odporność (ang. Robustness)
  - szum (noise),
  - Brakujące wartości (missing values),

  zdolność modelu do poprawnej predykcji klas w przypadku braku części danych lub występowania danych zaszumionych

Kryteria oceny metod klasyfikacji

- Zdolności wyjaśniania / Interpretowalność (ang. Interpretability): np. drzewa decyzyjne vs. sieci neuronowe

  odnosi się do stopnia w jakim konstrukcja klasyfikatora pozwala na zrozumienie mechanizmu klasyfikacji danych

- Skalowalność / Złożoność struktury (ang. Scalability), np.
  - rozmiar drzew decyzyjnego,
  - miary oceny reguły

  zdolność do konstrukcji klasyfikatora dla dowolnie dużych wolumenów danych

- Kryteria dziedzinowo zależne

Macierz pomyłek

- Analiza pomyłek w przydziale do różnych klas przy pomocy tzw. **Macierz pomyłek** (ang. *Confusion matrix*)
- Macierz $r \times r$, gdzie kolumny odpowiadają poprawnym klasom decyzyjnym, a wiersze decyzjom przewidywanym przez klasyfikator; na przecięciu wiersza *i* oraz kolumny *j* - liczba przykładów *n-ij* należących oryginalnie do klasy *i*-tej, a zaliczonej do klasy *j*-tej

- Przykład:

Przewidywane klasy decyzyjne	Rzeczywiste klasy		
	K_1	K_2	K_3
K_1	50	0	0
K_2	0	48	2
K_3	0	4	46

Ocena klasyfikatora binarnego

Niektóre problemy → jedna z klas posiada szczególne znaczenie, np. diagnozowanie poważnej choroby → **klasyfikacja binarna**.

	Rzeczywiste klasy	
Wynik klasyfikacji	Pozytywna (True)	Negatywna (False)
Pozytywna (True)	TP	FP (Błąd typu I)
Negatywna (False)	FN (Błąd typu II)	TN

Nazewnictwo (inspirowane medycznie):

- **TP** (ang. true positive) – liczba poprawnie sklasyfikowanych przykładów z wybranej klasy (ang. hit),
- **FP** (ang. false positive) – liczba przykładów błędnie przydzielonych do wybranej klasy, podczas gdy w rzeczywistości do niej nie należą (fałszywy alarm – z ang. *false alarm*).
- **FN** (ang. false negative) – liczba błędnie sklasyfikowanych przykładów z tej klasy, tj. decyzja negatywna podczas gdy przykład w rzeczywistości jest pozytywny (błąd pominięcia - z ang. miss),
- **TN** (ang. true negative) – liczba przykładów poprawnie nie przydzielonych do wybranej klasy (poprawnie odrzuconych z ang. correct rejection),

Miary klasyfikatora

- **Trafność klasyfikacji** (*ang. classification accuracy*) (*Acc*)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

- **Błąd klasyfikowania** (*ang. failed detection*) (*FD*)

$$FD = \frac{FN + FP}{TP + FN} \times 100\%$$

- **Wrażliwość / czułość** (*ang. sensitivity*) (*Se*)

$$SE = \frac{TP}{TP + FN} \times 100\%$$

- **Specyficzność** (*ang. specificity*) (*Sp*)

$$SP = \frac{TN}{FP + TN} \times 100\%$$

- Wnikliwszą analizę działania klasyfikatorów binarnych dokonuje się w oparciu o analizę krzywej ROC, (*ang. Receiver Operating Characteristic*).

Miary klasyfikatora na Macierzy pomyłek

		Reference Label			
		Abnormal	Normal		
Test Label	Abnormal	TP	FP		PPV
	Normal	FN	TN		NPV
		SE	SP		ACC

Analiza macierzy… spróbuj rozwiązać…

$$SE = \frac{TP}{TP + FN} \times 100\% = ???$$

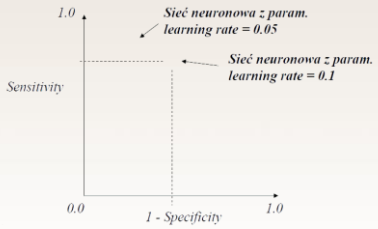
$$SP = \frac{TN}{FP + TN} \times 100\% = ???$$

Wynik klasyfikacji	Rzeczywiste klasy	
	1	0
1	60	80
0	30	20

- 60+30 = 90 przykładów w danych należało do Klasy 1
- 80+20 = 100 przykładów było w Klasy 0
- 90+100 = 190 łączna liczba przykładów

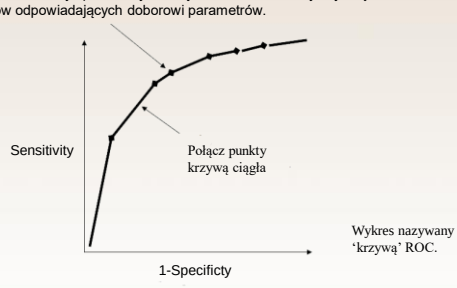
Analiza krzywej ROC

Każda technika budowy klasyfikatora może być scharakteryzowana poprzez pewne wartości miar 'sensitivity' i 'specificity'. Graficznie można je przedstawić na wykresie 'sensitivity' vs. 1 – 'specificity'.

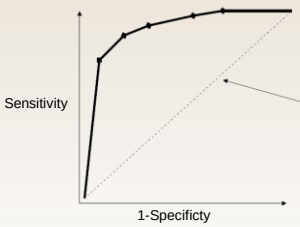


Krzywa ROC

Algorytm może być parametryzowany, i w rezultacie otrzymuje się serie punktów odpowiadających doborowi parametrów.



ROC - analiza



Im krzywa bardziej wygięta ku górnemu lewemu narożnikowi, tym lepszy klasyfikator .

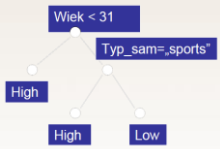
Przekątna odpowiada losowemu „zgadywaniu". Im bliżej niej, tym gorszy klasyfikator

Można porównywać działanie kilku klasyfikatorów.
Miary oceny np. AUC – pole pod krzywą...

Sformułowanie problemu

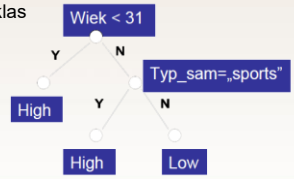
Dana jest baza danych przykładów, z których każdy należy do określonej klasy, zgodnie z wartością atrybutu decyzyjnego. Celem klasyfikacji jest znalezienie modelu dla każdej klasy

Wiek	Typ_sam	Ryzyko
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Klasyfikacja poprzez indukcje drzew decyzyjnych (1)

- drzewo decyzyjne jest grafem o strukturze drzewiastej, gdzie
  - każdy wierzchołek wewnętrzny reprezentuje test na atrybucie (atrybutach),
  - każdy łuk reprezentuje wynik testu,
  - każdy liść reprezentuje pojedyncza klasę lub rozkład wartości klas



Klasyfikacja poprzez indukcje drzew decyzyjnych (2)

- Drzewo decyzyjne rekurencyjnie dzieli zbiór treningowy na partycje do momentu, w którym każda partycja zawiera dane należące do jednej klasy, lub, gdy w ramach partycji dominują dane należące do jednej klasy
- Kady wierzchołek wewnętrzny drzewa zawiera tzw. **punkt podziału** (*ang. split point*)**,** którym jest test na atrybucie (atrybutach), który dzieli zbiór danych na partycje

Klasyfikacja poprzez indukcje drzew decyzyjnych (3)

- **Algorytm podstawowy**:

  algorytm zachłanny, który konstruuje rekurencyjnie drzewo decyzyjne metoda top-down w sposób „dziel i rządź" (ang. *divide-and-conquer*)
- Wiele wariantów algorytmu podstawowego.
- **Podstawowa różnica: kryterium p**odziału czyli sposobu w jaki tworzone są nowe węzły wewnętrzne w drzewie decyzyjnym, używanego podczas fazy budowania drzewa decyzyjnego
- Metoda podziału **powinna maksymalizować dokładność konstruowanego drzewa** decyzyjnego, lub innymi słowy minimalizować błędna klasyfikacje rekordów danych.

Fazy algorytmu

- Algorytm jest wykonywany w dwóch fazach:
  - **Faza 1:**

    Konstrukcja drzewa decyzyjnego w oparciu o zbiór treningowy
  - **Faza 2:**

    Obcinanie drzewa w celu poprawy dokładności, interpretowalności i uniezależnienia się od efektu przetrenowania

Konstrukcja drzewa

- W fazie konstrukcji drzewa, **zbiór treningowy jest dzielony na partycje**, rekurencyjnie, w punktach podziału do momentu, gdy każda z partycji jest „czysta" (zawiera dane należące wyłącznie do jednej klasy) lub liczba elementów partycji jest dostatecznie mała (spada poniżej pewnego zadanego progu)
- Postać testu stanowiącego punkt podziału zależy od kryterium podziału i typu danych atrybutu występującego w teście:

dla atrybutu ciągłego A, test ma postać wartość(A) < x, gdzie x należy do dziedziny atrybutu A, $x \in dom(A)$

dla atrybutu kategorycznego A, test ma postać wartość(A) ∈ X, gdzie $X \subset dom(A)$

Algorytm konstrukcji drzewa

```
Make Tree (Training Data D)
{
        Partition(D)
}
Partition(Data S)
{
        if (all points in S are in the same class)
        then
                return

        for each attribute A do
                evaluate splits on attribute A;

        use best split found to partition S into S1 and S2
        Partition(S1)
        Partition(S2)
}
```


Algorytm konstrukcji drzewa

- W trakcie budowy drzewa decyzyjnego, wybieramy taki atrybut i taki punkt podziału, określający wierzchołek wewnętrzny drzewa decyzyjnego, który „najlepiej" dzieli zbiór danych treningowych należących do tego wierzchołka
- Do oceny jakości punktu podziału zaproponowano szereg kryteriów (wskaźników)

Podsumowanie

- Metody klasyfikacji
- Kryteria oceny metod klasyfikacji
- Sformułowanie problemu
- Klasyfikacja poprzez indukcje drzew decyzyjnych