GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych

KLASYFIKACJA I REGRESJA cz. 3

Wojciech Waloszek

wowal@eti.pg.gda.pl

Teresa Zawadzka

tegra@eti.pg.gda.pl

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

GDAŃSK UNIVERSITY OF TECHNOLOGY

Naiwny klasyfikator Bayesa

- Naiwny klasyfikator Bayesa (NKB) wykorzystuje wprost pojęcie prawdopodobieństwa warunkowego,

- Jest to prosty klasyfikator, ale bardzo często można dzięki niemu uzyskać „zawstydzająco” dobre rezultaty predykcji

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zasada działania NKB

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

1. **Mamy wyróżniony atrybut decyzyjny, wyznaczający *klasy***
2. **Próbujemy stwierdzić, co o przynależności do konkretnych klas możemy powiedzieć na podstawie wartości konkretnych atrybutów.**
3. **Każdy atrybut oceniamy oddzielnie.**

GDAŃSK UNIVERSITY OF TECHNOLOGY

Klasyfikacja nowego przykładu

- Załóżmy, że chcemy sklasyfikować nowy przykład:

M	1200	30	wyższe	tak	?

- Zadajemy sobie pytania:
  - Co możemy powiedzieć o przynależności do złotych klientów na podstawie tego, że S.C. = M?
  - Co możemy powiedzieć o przynależności do złotych klientów na podstawie tego, że Wykształcenie = wyższe?
  - itd.

GDAŃSK UNIVERSITY OF TECHNOLOGY

S.C. = M?

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

Dla złotych klientów, klienci „małżeńscy” stanowią **2/4** populacji.

Dla nie-złotych klientów, klienci „małżeńscy” stanowią **1/5** populacji.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Wykształcenie = wyższe?

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

Dla złotych klientów, klienci z wykształceniem wyższym stanowią **2/4** populacji.

Dla nie-złotych klientów, klienci z wykształceniem wyższym stanowią **1/5** populacji.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Sam. = tak?

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

Dla złotych klientów, klienci z samochodem stanowią **3/4** populacji.

Dla nie-złotych klientów, klienci z samochodem stanowią **2/5** populacji.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Decyzja!

Do podjęcia decyzji potrzebujemy jeszcze względnej liczebności klas: **złoci klienci** stanowią **4/9** zbioru przykładów, a **nie-złoci klienci** stanowią **5/9** zbioru.

M	1200	30	wyższe	tak	?

Na „tak”: $2/4 \cdot 2/4 \cdot 3/4 \cdot \mathbf{4/9} \approx 0{,}083$

Na „nie”: $1/5 \cdot 1/5 \cdot 2/5 \cdot \mathbf{5/9} \approx 0{,}009$

Decyzja: **Z.K. = tak**

GDAŃSK UNIVERSITY OF TECHNOLOGY

Ale co za tym stoi?

Tak naprawdę, mówiąc formalnie, chcemy estymować prawdopodobieństwo dwóch zdarzeń:

P(Z.K = tak | *E*)
P(Z.K = nie | *E*)

Przy czym *E* jest złożonym zdarzeniem postaci:
S.C. = tak, Wykształcenie = wyższe, Sam. = tak

Decyzja zapada poprzez porównanie wartości dwóch estymatorów prawdopodobieństw.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Prawo Bayesa

Do zbudowania estymatorów wykorzystujemy tzw. prawo Bayesa:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Ponieważ interesuje nas jedynie porównanie wartości dwóch prawdopodobieństw skupiamy się na liczniku, porównując:

$$P(Z.K = tak|E) \cdot P(E) = P(E|Z.K = tak) \cdot P(Z.K = tak)$$

$$P(Z.K = nie|E) \cdot P(E) = P(E|Z.K = nie) \cdot P(Z.K = nie)$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Naiwność klasyfikatora Bayesa

Prawdopodobieństwo warunkowe zdarzenia złożonego E, stanowiącego złożenie ***niezależnych*** zdarzeń prostych E_1, E_2 … można wyrazić jako:

$$P(E|H) = P(E_1|H) \cdot P(E_2|H) \cdot \ldots$$

Stąd otrzymujemy ostatni element wzoru:

$$P(E|\text{Z.K} = \text{tak}) = P(\text{S.C.} = \text{M} \mid \text{Z.K} = \text{tak}) \cdot P(\text{Wykształcenie} = \text{wyższe} \mid \text{Z.K} = \text{tak}) \cdot P(\text{Sam.} = \text{tak} \mid \text{Z.K.} = \text{tak})$$

$$P(E|\text{Z.K} = \text{nie}) = P(\text{S.C.} = \text{M} \mid \text{Z.K} = \text{nie}) \cdot P(\text{Wykształcenie} = \text{wyższe} \mid \text{Z.K} = \text{nie}) \cdot P(\text{Sam.} = \text{tak} \mid \text{Z.K.} = \text{nie})$$

Przyjmujemy tu jednak „naiwnie”, że np. posiadanie samochodu nie ma nic wspólnego z posiadaniem wyższego wykształcenia.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Atrybuty numeryczne

- W klasycznej wersji NKB obsługuje jedynie atrybuty nominalne.
- Rozszerzenie klasyfikatora polega na doborze metody wyznaczenia estymatora P($A = x|H$) dla atrybutów numerycznych A,
- Można to zrobić poprzez przyjęcie konkretnej postaci rozkładu wartości danego atrybutu, najczęściej przyjmuje się rozkład normalny z estymowanymi parametrami rozkładu:

$$P(A = x|H) = \frac{1}{\sqrt{2\pi} \cdot \sigma_H} e^{-\frac{(x-\mu_H)^2}{2\sigma_H^2}}$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Parametry rozkładu

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

Dla złotych klientów, $\mu_{DOR} = 812{,}5$; $\sigma_{DOR} = 271{,}95$; $\mu_{Wiek} = 37{,}5$; $\sigma_{Wiek} = 5{,}57$.

Dla nie-złotych klientów, $\mu_{DOR} = 800$; $\sigma_{DOR} = 223{,}61$; $\mu_{Wiek} = 36{,}8$; $\sigma_{Wiek} = 17{,}66$.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Decyzja (2)!

Tym razem w decyzji uwzględniamy jeszcze wartości atrybutów numerycznych.

M	1200	30	wyższe	tak	?

Na „tak”: $2/4 \cdot 0{,}0005 \cdot 0{,}029 \cdot 2/4 \cdot 3/4 \cdot \mathbf{4/9} \approx 1{,}2 \cdot 10^{-6}$

Na „nie”: $1/5 \cdot 0{,}0003 \cdot 0{,}021 \cdot 1/5 \cdot 2/5 \cdot \mathbf{5/9} \approx 5{,}6 \cdot 10^{-8}$

Decyzja: **Z.K. = tak**

GDAŃSK UNIVERSITY OF TECHNOLOGY

Inne rozszerzenia NKB

- NKB w sposób naturalny obsługuje brakujące wartości atrybutów (usuwamy ze wzoru odpowiednie prawdopodobieństwo).
- NKB można rozszerzać poprzez dodawanie „drobnych” wartości prawdopodobieństw do poszczególnych estymatorów, co zapobiega estymacji zerowych prawdopodobieństw.
- Atrybuty numeryczne można obsługiwać, dobierając inne rozkłady, a także poprzez przeprowadzenie dyskretyzacji.

GDAŃSK UNIVERSITY OF TECHNOLOGY

NKB – podsumowanie

Naiwny klasyfikator Bayesa:

1. Zadanie: predykcja (klasyfikacja)
2. Struktura modelu: Bayesowski model prawdopodobieństw apriori
3. Funkcja oceny jakości: brak
4. Metody przeszukiwania: jednoprzebiegowa estymacja prawdopodobieństw
5. Dodatkowe założenia:

   Obsługa atrybutów numerycznych poprzez założenie ustalonego rozkładu

Regresja

- W przypadku przewidywania wartości numerycznych używamy pojęcia regresji,
- Najbardziej popularne metody predykcji numerycznej bazują właśnie na metodach regresji,
- Metody regresji mogą także być użyte do klasyfikacji

Regresja liniowa

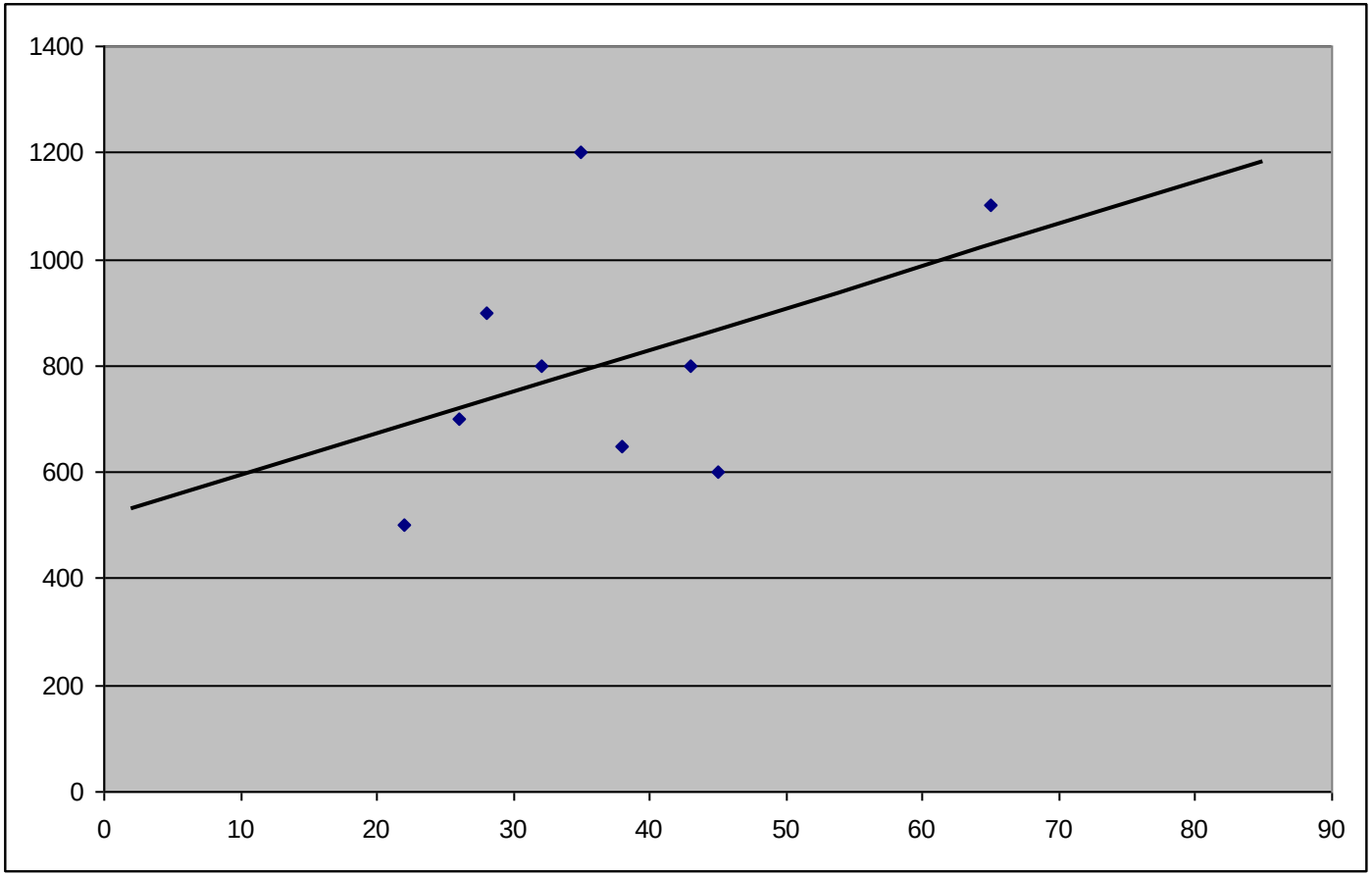
- Regresja liniowa polega na odnajdywaniu parametrów równania liniowego uzależniającego wartość jednego atrybutu od pozostałych:

$$x = w_0 + w_1 \cdot a_1 + w_2 \cdot a_2 + \ldots$$

- Wagi poszczególnych atrybutów (w_i) dobiera się, minimalizując pewną *funkcję błędu*, najczęściej tzw. błąd średniokwadratowy

GDAŃSK UNIVERSITY OF TECHNOLOGY

Wiek a D.O.R.



D.O.R. = 513 + 7,86 · Wiek

GDAŃSK UNIVERSITY OF TECHNOLOGY

Atrybuty nominalne a regresja

- Regresja liniowa w klasycznej postaci pozwala na obsługę wyłącznie atrybutów numerycznych.
- Atrybuty nominalne można zamienić na zbiór atrybutów ciągłych wg następującej recepty:
  - sortujemy wartości atrybutu *a* względem średniej wartości atrybutu przewidywanego,
  - tworzymy *n* – 1 atrybutów *a.i,* gdzie *n* to liczba możliwych wartości atrybutu nominalnego *a*,
  - atrybut *a.i* przyjmuje wartość 0 gdy *a* ma wartość najwyżej *i*-tą według przeprowadzonego sortowania

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przykład (wykształcenie)

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

Posortowane wartości, podstawowe (750), wyższe (766,67), średnie (862,5).

podstawowe → Wykształcenie-1 = 0, Wykształcenie-2 = 0

wyższe → Wykształcenie-1 = 1, Wykształcenie-2 = 0

średnie → Wykształcenie-1 = 1, Wykształcenie-2 = 1

GDAŃSK UNIVERSITY OF TECHNOLOGY

Regresja liniowa – podsumowanie

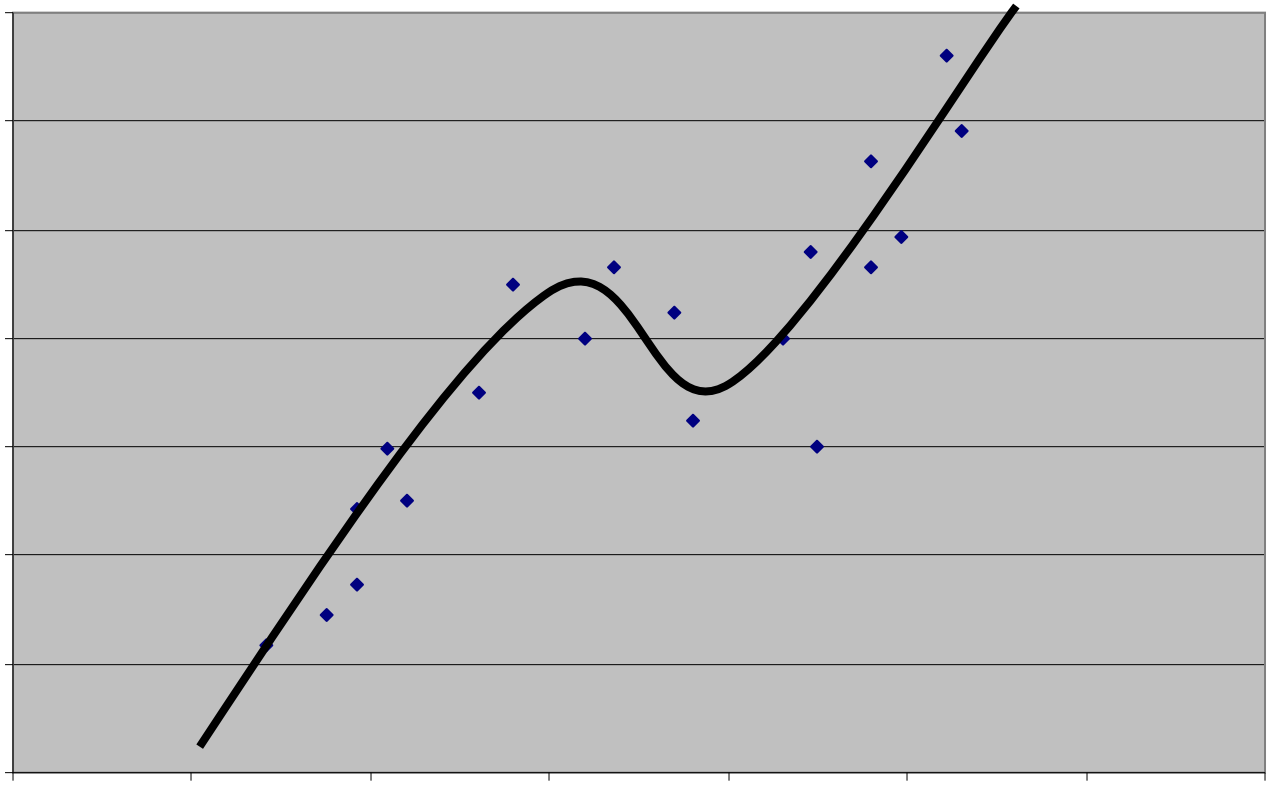
Regresja liniowa:

1. Zadanie: predykcja numeryczna
2. Struktura modelu: wektor wag atrybutów
3. Funkcja oceny jakości: funkcja błędu (średniokwadratowego)
4. Metody przeszukiwania: wyliczana bezpośrednio lub za pomocą dowolnej heurystycznej metody minimalizacji
5. Dodatkowe założenia:

   Obsługa atrybutów nominalnych poprzez zamianę ich na serię atrybutów numerycznych

GDAŃSK UNIVERSITY OF TECHNOLOGY

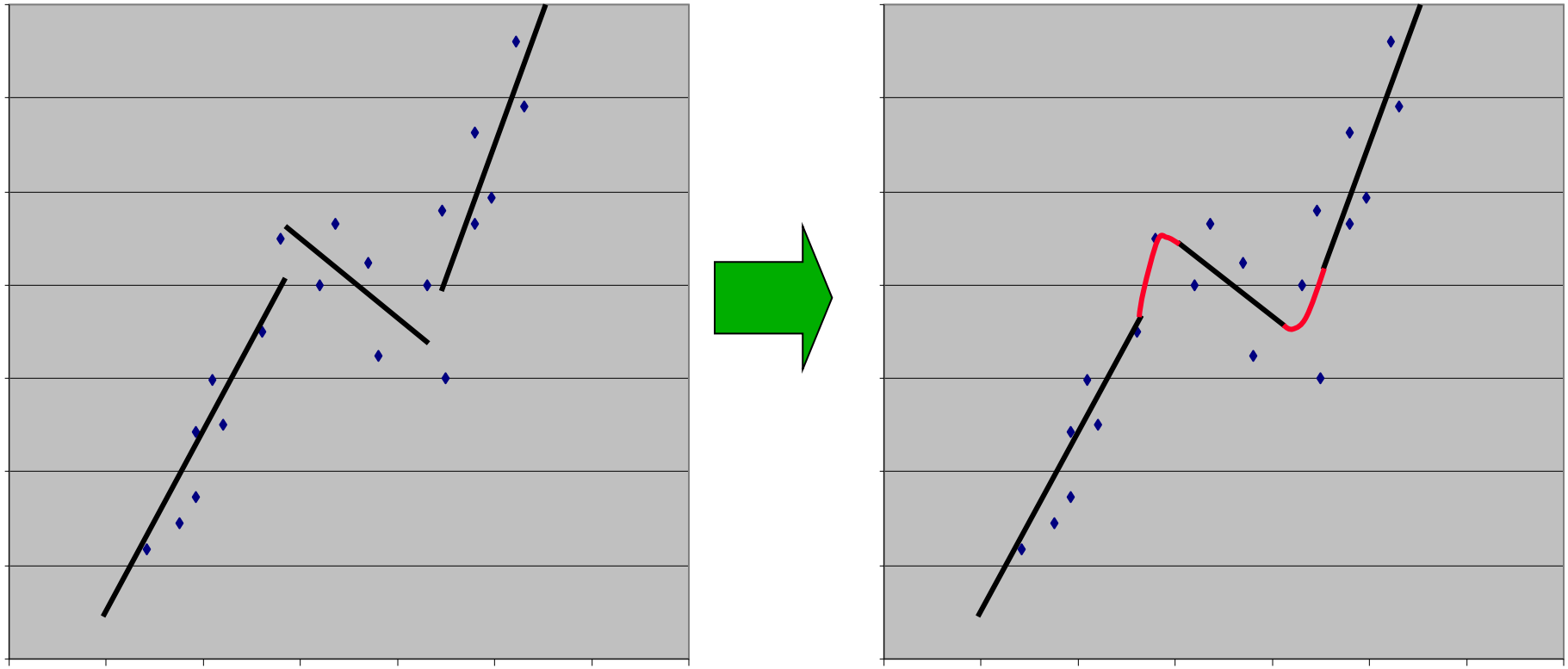
Rozszerzone metody regresji – wyższe stopnie



Wielomian trzeciego stopnia

GDAŃSK UNIVERSITY OF TECHNOLOGY

Rozszerzone metody regresji – fragmenty



Segmentowa regresja liniowa z wygładzaniem

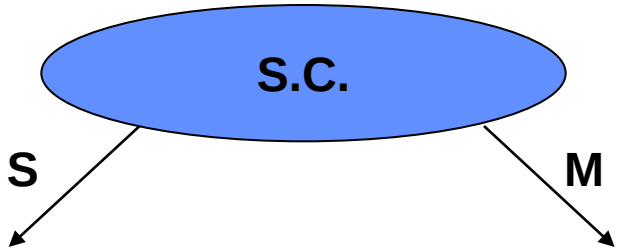
GDAŃSK UNIVERSITY OF TECHNOLOGY

Drzewa predykcji numerycznej

- Drzewa predykcji numerycznej rozwijają pomysł regresji liniowej dokonywanej segmentami,
- Drzewa dokonują podziału przykładów względem wartości pozostałych atrybutów, tak aby w liściach pozostawały zbliżone wartości atrybutu, którego wartość przewidujemy,
- Drzewa mogą mieć postać *drzewa regresji* (gdy w liściach mamy parametry rozkładu wartości) lub *drzewa modelu*, gdy w liściach zawarte są modele regresji liniowej

GDAŃSK UNIVERSITY OF TECHNOLOGY

Drzewo modelu



D.O.R. = 340.28 * S.C.=S +
+ 187.5 * Wyksztalcenie=srednie +
+ 458.33

D.O.R. = 291.67 * S.C.=S +
+ 260.71 * Wyksztalcenie=srednie +
+ 538.09

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dobór najlepszego atrybutu

- Dobór atrybutu odbywa się na zasadzie minimalizacji odchylenia standardowego:

$$SDR = \sigma(P) - \Sigma_i \left((|P^i|/|P|) \cdot \sigma(P^i) \right)$$

- Testy dla wartości numerycznych są binarne z wartością minimalizującą SDR,
- Atrybuty nominalne obsługiwane są zgodnie ze sposobem opisanym wcześniej dla regresji

GDAŃSK UNIVERSITY OF TECHNOLOGY

Inne rozszerzenia algorytmu

- Brakujące wartości atrybutów obsługiwane są poprzez zastępowanie ich wartością średnią lub poprzez „zgadywanie” wartości atrybutu na podstawie wartości atrybutu najbardziej skorelowanego,

- „Wygładzanie” osiąga się poprzez przechowywanie wartości w każdym węźle drzewa (nie tylko w liściach) oraz poprzez uśrednianie

Drzewa predykcji numerycznej – podsumowanie

Drzewa predykcji numerycznej:

1. Zadanie: predykcja numeryczna
2. Struktura modelu: *drzewo regresji* lub *drzewo modelu*
3. Funkcja oceny jakości: minimalizacja odchylenia standardowego
4. Metody przeszukiwania: zachłanna, divide-and-conquer
5. Dodatkowe założenia:

   Obsługa atrybutów nominalnych poprzez zamianę ich na serię atrybutów numerycznych

   Obsługa wartości brakujących poprzez dobór atrybutu najbardziej skorelowanego

   Przycinanie metodą walidacji krzyżowej

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dziękujemy za uwagę

Zapraszamy na wykład:

OCENA KLASYFIKATORÓW