Data Mining — techniki, algorytmy i zastosowania

18 listopada Warszawa

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
Warszawska Wyższa Szkoła Informatyki
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Marcin Szeliga

- Piętnastoletnie doświadczenie w pracy z serwerem SQL
- Trener i konsultant
- Autor książek i artykułów
- Microsoft Most Valuable Professional w kategorii SQL
- Specjalista technologii Microsoft

MVP Microsoft® Most Valuable Professional

POLISH SQL SERVER USER GROUP

śląskaregionalna grupamicrosoft

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI

Warszawska Wyższa Szkoła Informatyki

UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

O czym będzie ta sesja?

- **Wprowadzenie**
- Omówienie procesu eksploracji danych
- Terminologia
- Algorytmy eksploracji danych

Eksploracja danych

- Technologie do analizowania danych i wykrywania (bardzo) ukrytych układów
- Dość nowe (<20 lat), ale skuteczne algorytmy opracowane na drodze badań nad bazami danych
- Połączenie statystyki, analizy prawdopodobieństwa i technologii baz danych

Funkcje eksploracji danych

Badanie danych

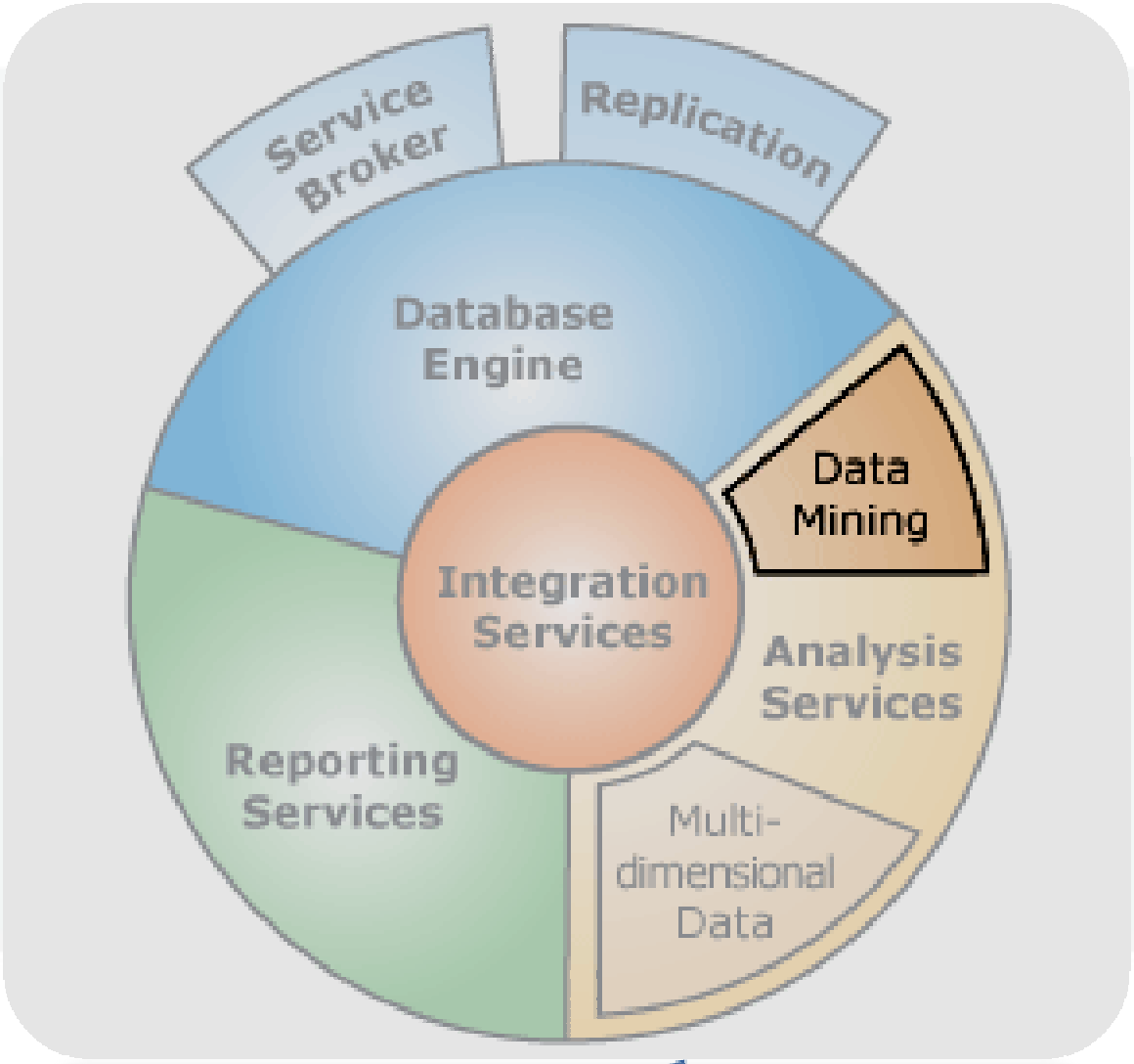
Szukanie układów

Sporządzanie prognoz

Eksploracja vs analiza danych

Zastosowania eksploracji danych

DM – część Microsoft SQL Server

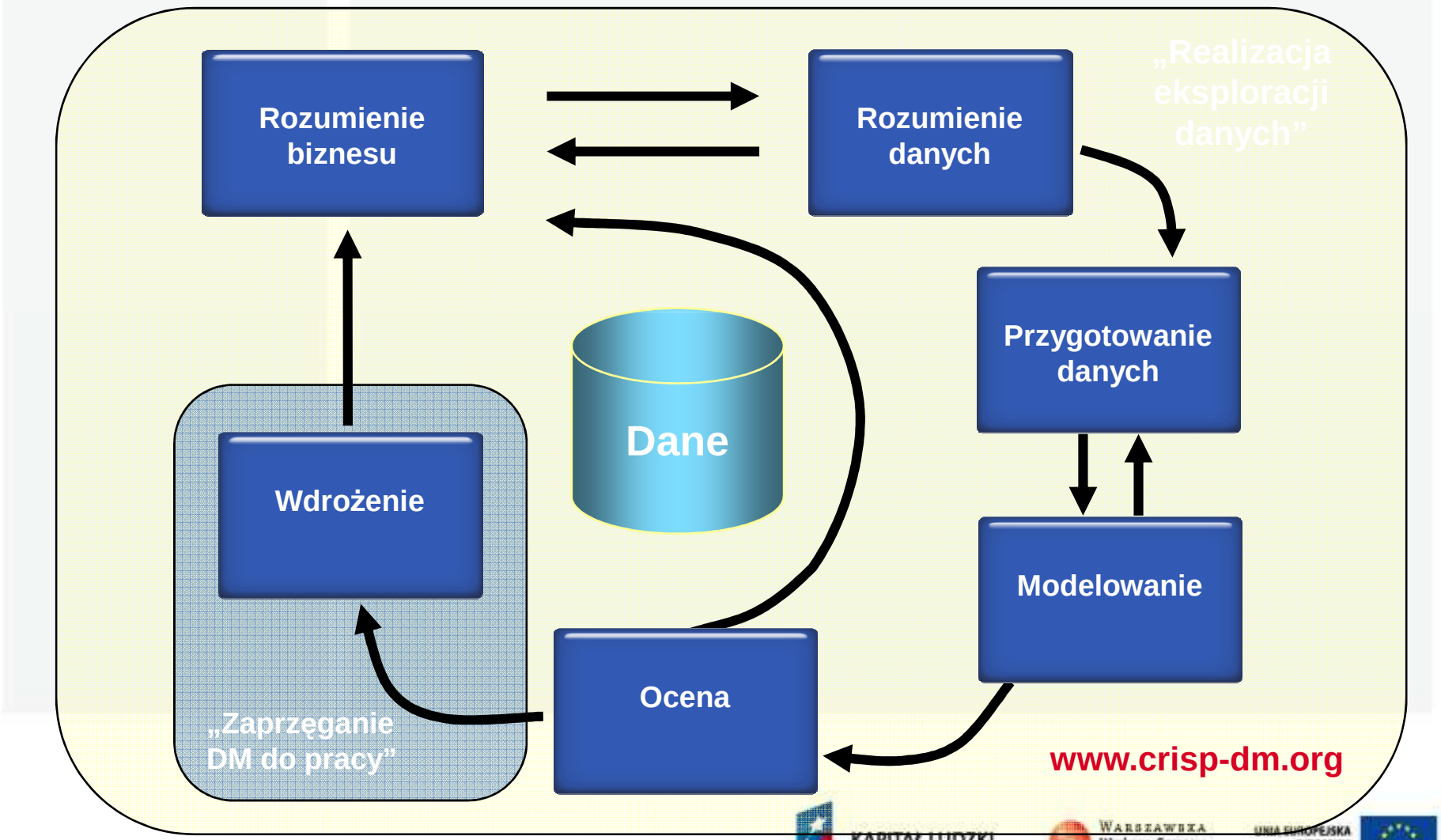


Demo

- Kto miał szansę uratować się z katastrofy?
- Kiedy warto wziąć kredyt walutowy?
- Kto podał błędne dane?

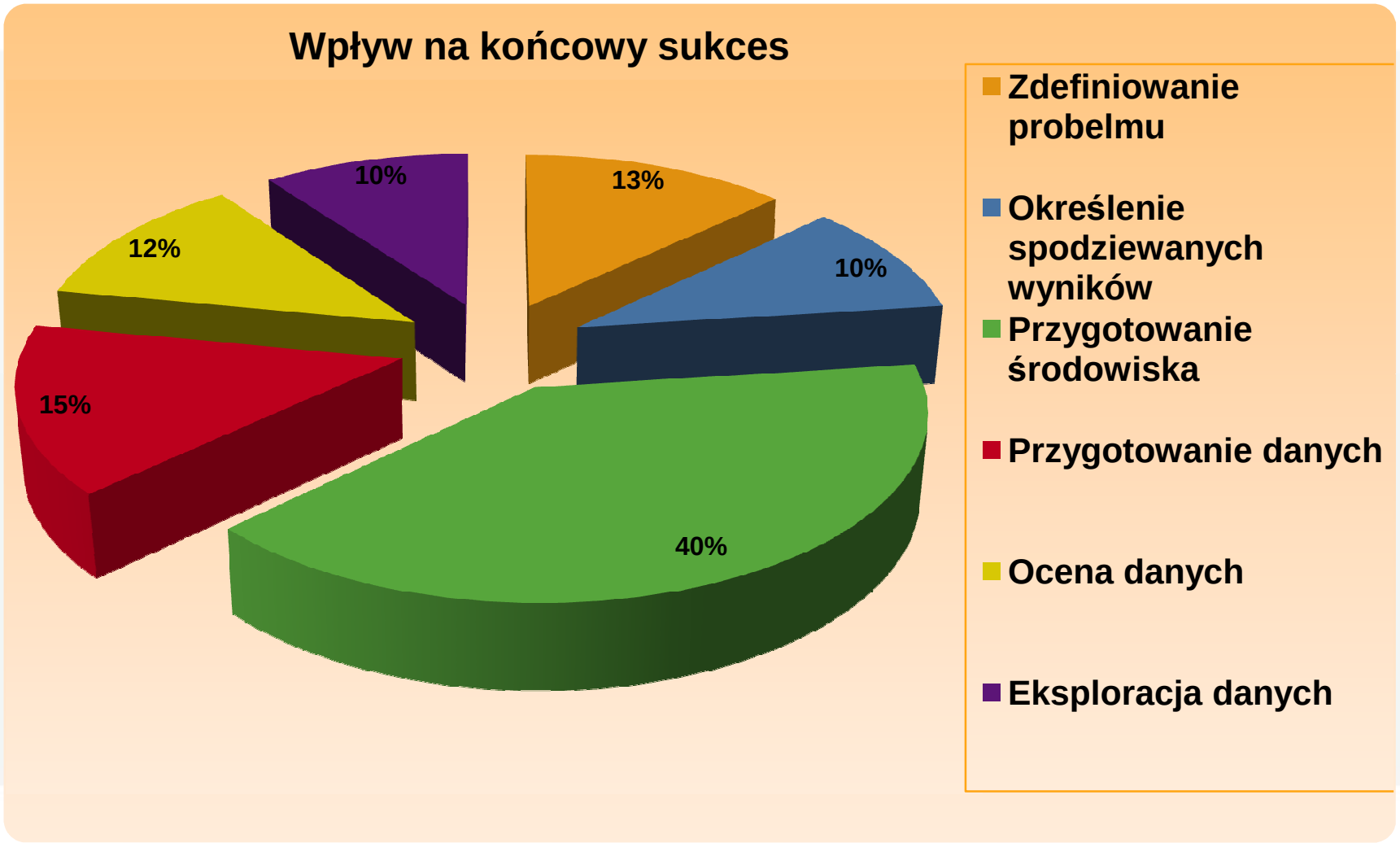
Proces eksploracji danych

CRISP-DM

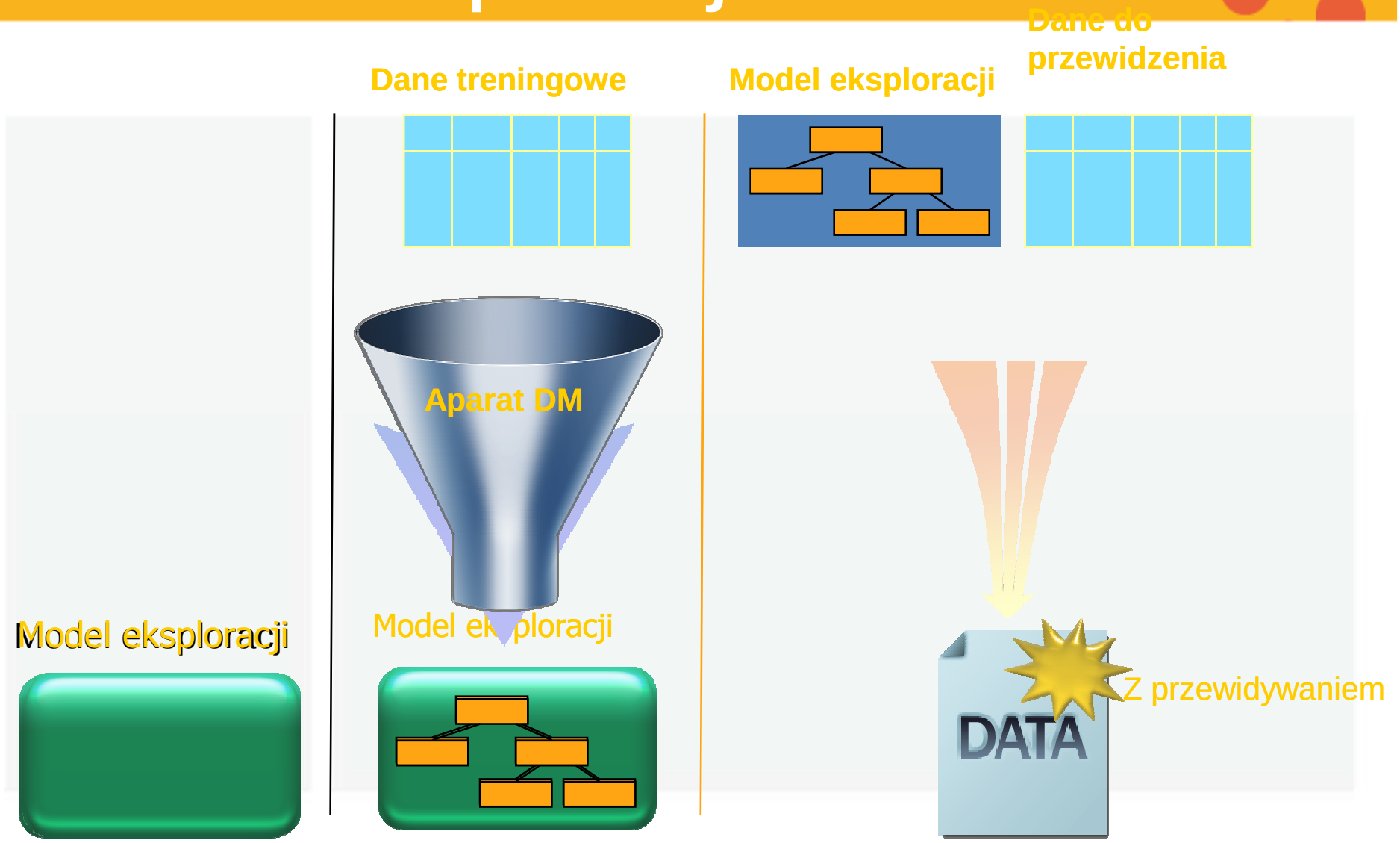


Proces eksploracji danych

Proces eksploracji danych



Proces eksploracji



Etapy procesu życia modelu DM

1. Definicja modelu
   - Definiowanie kolumn dla przypadków: wizualnie (BIDS, Excel), przy użyciu DMX lub z języka PMML
2. Trening modelu
   - Wprowadzenie dużej ilości danych z rzeczywistej BD lub z dziennika systemu
3. Testowanie modelu
   - Dane testowe muszą być inne niż treningowe
4. Używanie modelu (eksploracja i przewidywanie)
   - Używanie modelu na nowych danych w celu przewidywania wyników
5. Aktualizacja modelu
   - Co miesiąc, co tydzień, co noc lub częściej – ponowne testowanie

Demo

- Przygotowanie danych treningowych ma znaczenie

Struktura eksploracji (Mining Structure)

- Opisuje dane, którymi trzeba się zająć
  - Kolumny ze źródeł danych i ich:
    - Typy danych
    - Typy zawartości
- Zawiera modele eksploracji
  - Często buduje się kilka różnych modeli w jednej strukturze
- Przechowuje dane treningowe, zwane przypadkami (cases) (jeśli to konieczne)
- Przechowuje dane testowe, określane jako wydzielone (holdout) (w programie SQL 2008)

Model eksploracji danych (Data Mining Model)

- Kontener układów odkrywanych za pomocą algorytmu eksploracji danych w przypadkach treningowych
  - Tabela zawierająca układy
    - Wyrażone przez wizualizatory
- Określa użycie kolumn już zdefiniowanych w strukturze eksploracji

Przypadki: to, co badamy (Cases)

- **Przypadek** – zbiór kolumn (atrybutów), które mają być analizowane
  - Wiek, płeć, region, roczne wydatki
- **Klucz przypadku** – unikatowy identyfikator przypadku
- **Atrybut** – cecha przypadku (im więcej atrybutów tym więcej potrzebować będziemy przypadków)
- **Stan** – wartość atrybutu (liczba stanów najczęściej jest ograniczana). Wszystkie atrybuty mają specjalny stan Missing
- Kolumna ma:
  - Typ danych
  - Typ zawartości
  - I opcjonalnie dystrybucję, dyskretyzację, pokrewne kolumny, opcje (np. NOT NULL)

Typy danych kolumny (Column Data Types)

- Nie zajmujemy się typami niskiego poziomu
- Typy używane w eksploracji danych:
  - **Text**
  - **Long**
  - **Boolean**
  - **Double**
  - **Date**
  - Plus specyficzne dla niektórych algorytmów firm trzecich jak:
    - Time i Sequence

Typy zawartości kolumny (Content Types)

Sterują algorytmami

- Typowe:
  - DISCRETE
    - Czerwony, Niebieski, Zielony
  - CONTINOUS
    - 6511,49 €
  - DISCRETIZED
    - 1-5, 6-20, 21+
- Oznacza klucz:
  - KEY
- Do celów specjalnych:
  - KEY SEQUENCE
  - KEY TIME
  - ORDERED
  - CYCLICAL

Użycie kolumn (Column Usage)

- W niektórych algorytmach interpretacja tego nieznacznie się różni, ale na ogół kolumny służą do:
  - INPUT
    - W celu przewidywania innej kolumny
  - PREDICT
    - Te kolumny są zarówno przewidywane, jak i działają jako wejściowe do przewidywania innych
  - PREDICT_ONLY
    - Nieużywane jako dane wejściowe
- Wszystkie kolumny mogą być wejściowymi i przewidywanymi

Dystrybucja wartości

- Jeśli jest znana dystrybucja wartości (należy ją znać), trzeba ją podać:
  - NORMAL
    - Typowa krzywa Gaussa (dzwonowa)
  - LOG NORMAL
    - Większość wartości na „początku” skali
  - UNIFORM
    - Linia płaska – jednakowo prawdopodobna lub idealnie losowa
- Mogą istnieć inne dystrybucje, ale nie można ich podać – algorytm będzie działać prawidłowo

Dyskretyzacja

- Bardzo ważna technika
- Gdy nie ma potrzeby analizowania pełnego ciągłego zakresu
- Eksploracja danych może automatycznie konwertować dane na przedziały
  - Domyślnie na 5
- Techniki:
  - AUTOMATIC
  - CLUSTERS
  - EQUAL_AREAS

I wreszcie

- Przypadek zagnieżdżony (Nested Case) – przypadek zawierający tabelę kolumn
  - Zakupy klienta
- Używany do analizy układów w relacji

- Ma zagnieżdżony klucz (Nested Key)
  - Nie „relacyjny” klucz obcy!
  - Na ogół klucz zagnieżdżony to kolumna, która ma być analizowana
    - Np.: Nazwa produktu lub model

Przypadek zagnieżdżony

KlientID	Płeć	Stan cywilny	Wykształcenie	Własność domu	Mebel
980001	M	ZamŻon	Licencjaci	Wynajem	Sofa
					Telewizor
980002	M	ZamŻon	Licencjaci	Własność	Drabina
					Boiler
					Sofa
980003	K	Wolny	Magistrowie	Własność	Leżanka
980004	M	Wolny	Średnie	Własność	Boiler
					Telewizor
					Odtwarzacz DVD
980005	K	ZamŻon	Licencjaci	Wynajem	Stelaż
					Telewizor
980006	K	ZamŻon	Magistrowie	Wynajem	Regał na książki
					Mata do jogi
					Waza

Trening

- Użycie instrukcji INSERT INTO
  - Wprowadza przypadki do aparatu
- Użycie składni SHAPE do tworzenia zagnieżdżonych zestawów wierszy wejściowych

- Ważne:
  - Używać jedynie danych treningowych (na ogół ok. 70%)
  - Pozostawić z boku trochę danych testowych

Ile treningu?

- Brak sztywnych reguł odnośnie do liczby przypadków
- Niemożliwe przetrenowanie przez podanie zbyt wielu przypadków
  - Możliwe przetrenowanie w źle sparametryzowanych modelach
    - Zbyt szczegółowe modele o za małym stopniu ogólności
- Czy są używane reprezentatywne próbki?
  - Nie jest potrzebna duża ilość danych treningowych
- Trening jest wystarczający, gdy walidacja modelu jest poprawna (patrz później)

Testowanie i walidacja

- Sprawdzenie poprawności modelu
  - Dokładność
    - Czy zapewnia poprawne korelacje i przewidywania?
  - Wiarygodność
    - Czy działa podobnie w odniesieniu do innych danych testowych?
  - Przydatność
    - Czy zapewnia wgląd w dane czy tylko oczywistości?

Walidacja modelu

- Typowe podejścia:
  - Dokładność
    - Wykresy wzrostu i zysku
    - Wykresy punktowe
    - Macierz klasyfikacji
  - Wiarygodność
    - Walidacja krzyżowa
    - Przewidywania danych zewnętrznych
  - Przydatność
    - Wymaga przejrzenia przez eksperta z konkretnej dziedziny
      - Może wystarczyć proste sprawdzenie korelacji atrybutów

Zautomatyzowane testowanie

- *Znakomita* funkcja DM programu SQL Server
- Klikniecie karty „Mining Accuracy” automatycznie i szybko wykonuje test:
  - Do przewidywania wartości są używane dane testowe
  - Wyniki tego przewidywania są porównywane ze znaną wartością (w wydzieleniu)
  - Wyniki to:
    - Wykres wzrostu, wykres zysku, wykres punktowy, macierz klasyfikacji, statystyki walidacji krzyżowej

Przewidywanie!

- Zastosowanie modelu do przewidywania nieznanych danych
- Użycie instrukcji SELECT
  - Klauzula PREDICTION JOIN
- Zwrócone wartości mogą zawierać tabele
  - Na zagnieżdżonych tabelach można wykonywać dalsze instrukcje SELECT

Funkcje

- Funkcji DMX można używać do tworzenia bardziej rozbudowanych wyrażeń przewidywań
- Przewidywanie miar statystycznych:
  - `PredictProbability`
  - `PredictHistogram`
- Użycie ma kluczowe znaczenie podczas przewidywania dowolnych wartości, w szczególności zysku lub ryzyka

PredictProbability

PredictProbability(LoanStatus)	Prawdopodobieństwo najbardziej prawdopodobnego wyniku
PredictProbability(LoanStatus, “Defaulted”)	Prawdopodobieństwo, że pożyczka będzie bardzo kłopotliwa

- Podobnie jak PredictAdjustedProbability itp.

Nie zapominać o eksploracji i analizie

- Kilka znakomitych wizualizatorów firmy Microsoft
- Dostępne w: BIDS, SSMS, SSRS, Excel, Visio oraz
  w wersji dla aplikacji użytkownika
- Wiele z nich będzie można obejrzeć dziś po południu!

- Wyszukując układy, można też wykonywać kwerendy bezpośrednio na modelu eksploracji
  - Wiele przykładów można znaleźć w witrynie SQL Books Online. Można też postarać się o książkę „Data Mining with SQL Server 2008” (autorzy Jamie McLennan i ZhaoHui Tang)

Demo

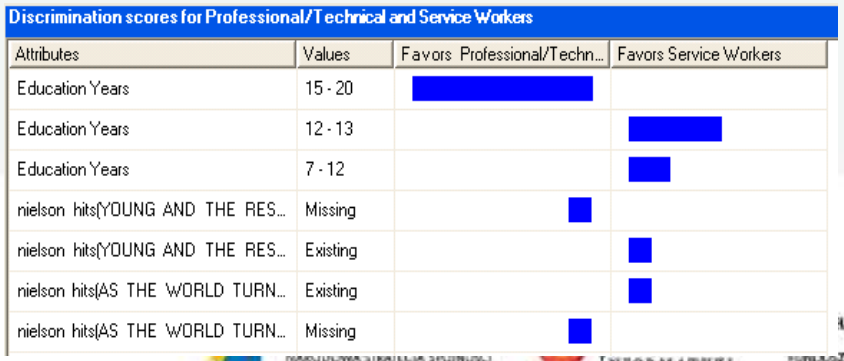
- Jaki film chcesz obejrzeć?
- Klasyfikacja potencjalnych klientów

Dostępne algorytmy

- **Naiwny klasyfikator Bayesa**
- **Drzewa decyzyjne**
- **Szeregi czasowe**
- **Klastrowanie**
- **Klastrowanie sekwencyjne**
- **Reguły asocjacyjne**
- **Regresja logistyczna**
- **Sieci neuronowe**
- **Regresja liniowa**

Microsoft Naive Bayes

- Klasyczny algorytm uczenia przez obserwację
  - Skoro Kowalski spóźniał się codziennie przez 5 lat, to jutro pewnie też się spóźni
- Należy do klasyfikatorów liniowych i nie nadaje się do rozwiązywania nieliniowych problemów
  - Takich jak określenie koloru pola szachownicy



Microsoft Naive Bayes

- Połączenie prawdopodobieństwa warunkowego i bezwarunkowego
  - Bezwarunkowe (początkowe) prawdopodobieństwo zależy od rozkładu przypadków
    - 60 % klientów to kobiety
  - Warunkowe prawdopodobieństwo zależy od zaobserwowanych faktów
    - 80% mężczyzn dzwoni dwa razy

Microsoft Naive Bayes

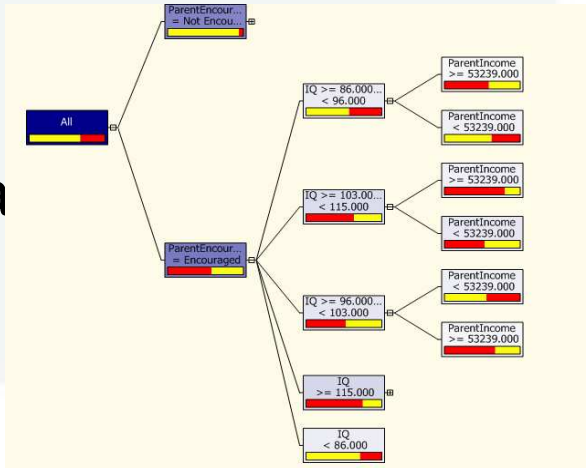
- Prawdopodobieństwa są mnożone:
  - Każde z nich ma taki sam wpływ na wynik …
  - O ile atrybuty wejściowe są od siebie niezależne
- Ilość zwracanych korelacji można kontrolować
  - MINIMUM_DEPENDENCY_PROBABILITY

Microsoft Naive Bayes - podsumowanie

- Wymaga starannego przygotowania danych
  - Wyłącznie wartości dyskretne
  - W praktyce wiele atrybutów jest ze sobą powiązanych
- Szybki
  - Nawet dla wielu przewidywanych atrybutów
- Niedokładne i mało wiarygodne predykcje
- Używany głownie do klasyfikacji tekstów oraz na etapie przygotowywania danych

Microsoft Decision Trees

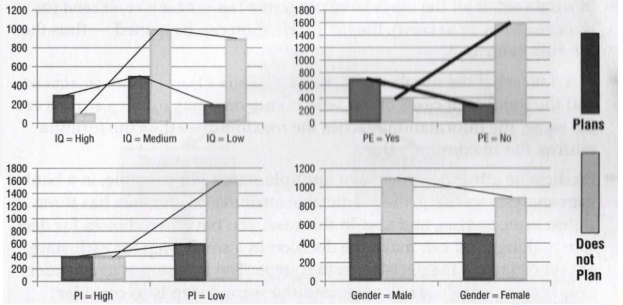
- Klasyczny algorytm wspomagania procesu decyzyjnego
- Dla każdego przewidywanego atrybutu tworzone jest osobne drzewo
- W pierwszej kolejności obliczane są zależności pomiędzy atrybutami wejściowymi
  - Ich miarą jest poziom entropii, punkty Bayesa lub ważone punkty Bayesa
    - Jeżeli wszyscy planują studia, wynikiem jest 0
    - Jeżeli połowa osób planuje studia wynik jest maksymalny
    - SCORE_METHOD



Microsoft Decision Trees

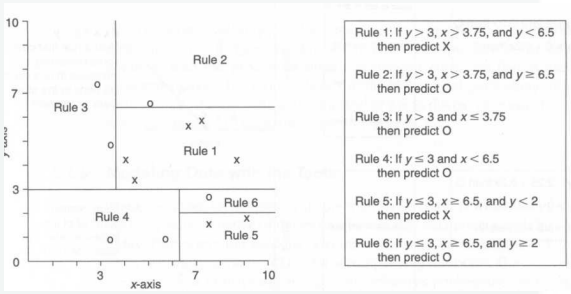
- Wybierany jest atrybut którego podział wyznaczy korzeń drzewa
- Następnie przeprowadzany jest rekurencyjny podział danych na podzbiory
  - Każdy atrybut wejściowy ma wpływ na podział
    - COMPLEXITY_PENALTY
- Jeżeli przewidywany atrybut jest ciągły, przeprowadzana jest regresja
  - Reguła regresji będzie inna dla każdego węzła drzewa
    - FORCE_REGRESSOR

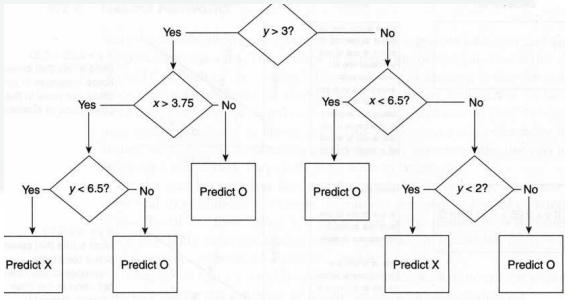
		IQ			Parent Encouragement		Parent Income		Gender	
		High	Medium	Low	Yes	No	High	Low	Male	Female
College Plan	Plans	300	500	200	700	300	400	600	500	500
	Does not Plan	100	1000	900	400	1600	400	1600	1100	900



Microsoft Decision Trees

- Wartości mogą być dzielone binarnie lub całkowicie
  - ryzyko = wysokie lub ryzyko = inne niż wysokie
  - ryzyko = wysokie lub ryzyko = średnie lub ryzyko = niskie
  - SPLIT_METHOD
- Wynikiem jest zbiór reguł dzielących różnorodny zbiór wejściowy na jednorodne (względem przewidywanego atrybutu) podzbiory
  - Intuicyjne
  - Łatwe do przekształcenia w klauzulę WHERE



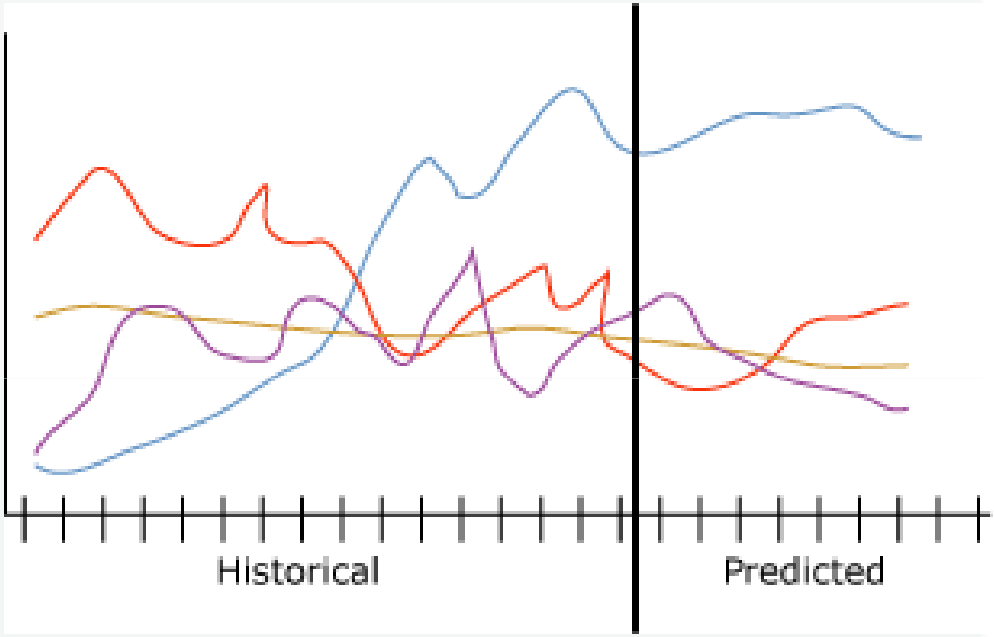


Microsoft Decision Trees - podsumowanie

- Nie najlepiej przewiduje ciągłe wartości
- Dynamicznie grupuje stany atrybutów wejściowych
  - Atrybuty ciągłe poddawane są dyskretyzacji
    - Algorytm grupuje je dzieląc na takie same przedziały
    - Następnie przedziały są łączone, jeżeli zwiększy to szansę podziału
    - Do podziału wybierany jest najwyżej oceniony przedział
- Zachłanna metoda „dziel i zwyciężaj”
  - Jeżeli wartości pewnych atrybutów są ze sobą silnie powiązane (np. wykształcenie i dochód) tylko jeden z nich zostanie użyty do podziału drzewa
- Możliwość przetrenowania
- Zbyt wiele możliwych wartości atrybutu (np. kody pocztowe) – algorytm uwzględnia tylko 99 najpopularniejszych wartości + specjalną wartość Missing
- Zwraca intuicyjne i łatwe do zastosowania wyniki

Microsoft Time Series

- Zastosowanie drzew regresji do opisywania i przewidywania wartości szeregu
  - Drzewa umożliwiają stosowanie wielu regresorów

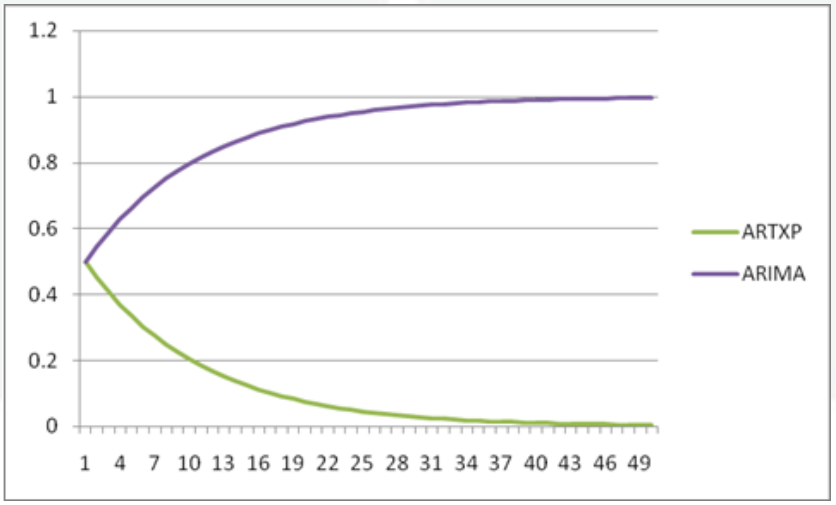


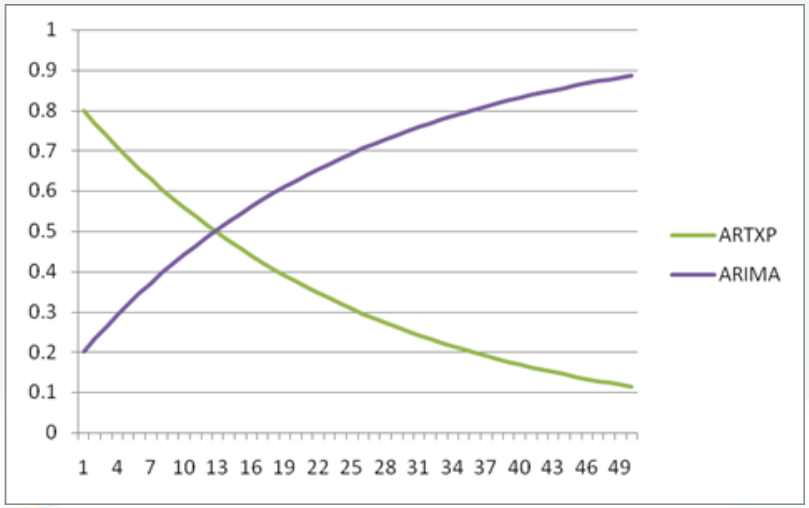
Microsoft Time Series

- Dane są bardzo sezonowe
  - Sezonowość wykrywana za pomocą szybkiej transformacji Fouriera
    - AUTO_DETECT_PERIODICITY
    - PERIODICITY_HINT
- Szeregi czasowe
  - W programie SQL Server 2005 jest używany algorytm ARTXP (drzewa autoregresyjne z predykcją krzyżową)
    - Do prognozowania krótkoterminowego
  - W programie SQL Server 2008 jest używana hybryda poprawionego algorytmu ARTXP standardowego algorytmu ARIMA (scałkowana autoregresja i średnia ruchoma)
    - Znakomite do prognozowania krótko- i długoterminowego

Microsoft Time Series

- ARIMA jest klasycznym algorytmem predykcyjnym
  - Na podstawie historycznych danych i różnic między nimi przewiduje przyszłe dane
- Scalanie wyników ARTXP i ARIMA
  - PREDICTION_SMOOTHING





PREDICTION_SMOOTHING = 0.5

PREDICTION_SMOOTHING = 0.2

Microsoft Time Series

Miesiąc	Mleko	Chleb
Sty	100	80
Lut	120	90
Mar	110	85
Kwi	115	110
Maj	125	120
Cze	120	123
Lip	140	150
…		

Format A

Miesiąc	Produkt	Sprzedaż
Sty	Mleko	100
Sty	Chleb	80
Lut	Mleko	120
Lut	Chleb	90
Mar	Mleko	110
Mar	Chleb	85
Kwi	Mleko	115
…		

Format B

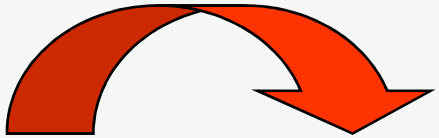
Microsoft Time Series

- Przykładowe zastosowania:
  - Prognozowanie
- Algorytm został tak skonfigurowany, że zwraca jak najlepsze wyniki przy minimalnej liczbie danych wejściowych
- Specyficzne wymagania:
  - Dane muszą być kompletne (choć serie mogą zacząć się w dowolnym punkcie czasu)
    - MISSING_VALUE_SUBSITUTION
  - Dane muszą być posortowane według klucza czasu
  - Przewidywany atrybut musi być ciągły

Regresja i autoregresja

- Regresja polega na wyznaczeniu wartości atrybuty X na podstawie wartości innych atrybutów
- Autoregresja – na podstawie wcześniejszych wartości atrybutu

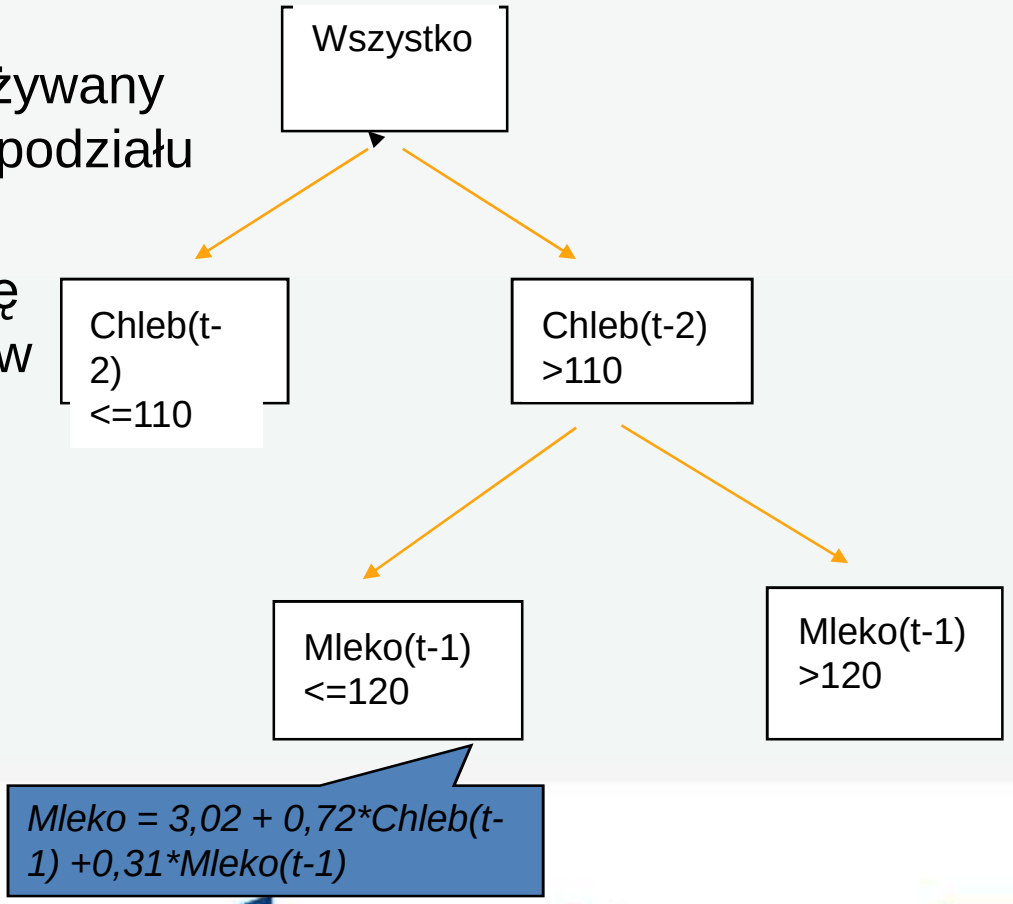
Miesiąc	Mleko	Chleb
Sty	100	80
Lut	120	90
Mar	110	85
Kwi	115	110
Maj	125	120
Cze	120	123
Lip	140	150
…		



Id Przyp	Mleko (t-2)	Mleko (t-1)	Mleko (t0)	Chleb (t-2)	Chleb (t-1)	Chleb (t0)
1	100	120	110	80	90	85
2	120	110	115	90	85	110
3	110	115	125	85	110	120
4	115	125	120	110	120	123
5	125	120	140	120	123	150
…						

Drzewo autoregresji

- Chociaż klucz czasu jest używany jako atrybut wejściowy (a więc jest uwzględniany przy tworzeniu węzłów), drzewo regresji nie oddaje sezonowości danych
- Każdy obliczony węzeł jest używany do wyznaczenia następnego podziału
- Ignorując niektóre formuły regresji ARTXP nie nadaje się do prognozowania >10 kroków

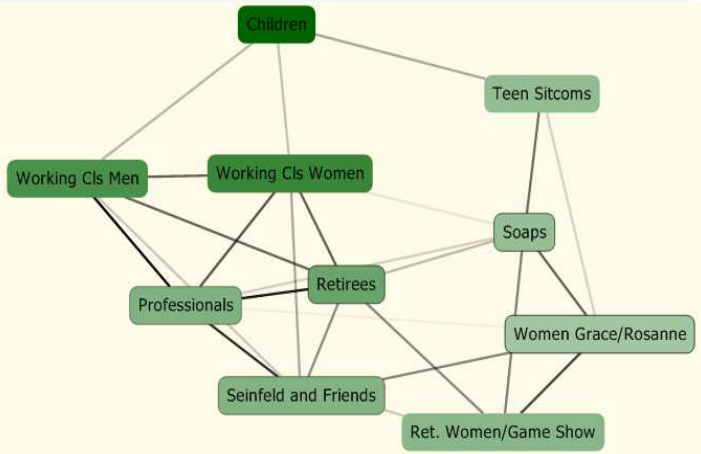


Microsoft Time Series - podsumowanie

- Wymaga pojedynczego wskaźnika czasu
- Niezgodny ze standardem Predictive Model Markup Language
- Pozwala wybrać stosowany algorytm
  - FORECAST_METHOD
- Umożliwia ocenę predykcji
  - HISTORIC_MODEL_COUNT
  - HISTORICAL_MODEL_GAP
- Automatycznie wykrywa mało wiarygodne predykcje
  - INSTABILITY_SENSITIVITY

Microsoft Clustering

- Segmentacja danych na podstawie ukrytych w nich zależności
  - W obrębie segmentów różnice pomiędzy przypadkami powinny być jak najmniejsze, pomiędzy segmentami – jak największe
- Proces heurystyczny
  - Podajemy liczbę klastrów
  - Pierwszy układ klastrów jest pseudolosowy (we wszystkich wymiarach)
    - Ponieważ od niego w dużym stopniu zależą wyniki, algorytm tworzy kilka początkowych układów kandydujących i na końcu wybiera najlepszy
  - Następnie zmienia się klastry tak, aby przypadki znalazły się w ich środkach
  - Proces powtarza się dopóki przypadki zmieniają klastry
    - MODELLING_CARDINALITY



Microsoft Clustering – techniki segmentacji

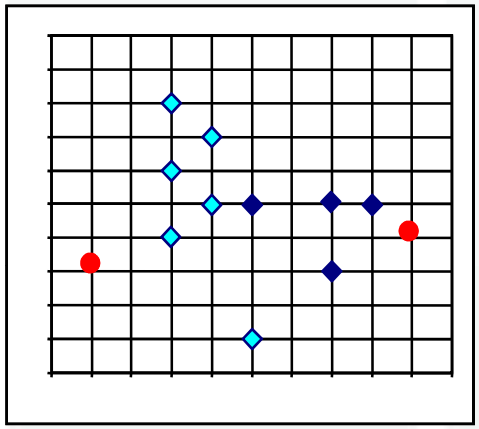
- K-Means
  - Klastry są reprezentowane przez ich centralne punkty
  - Obliczana jest odległość Euklidesa przypadków od klastrów

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

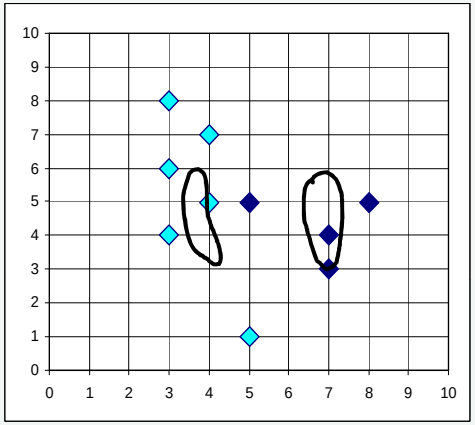
  - W przypadku atrybutów dyskretnych ocena odległości jest arbitralna
  - Dokładna tylko dla sferycznych manifoldów
    - Bardzo czuły na skrajne wartości
  - Przypadek trafi tylko do jednego (najbliższego) klastra
  - Odległości nie są bezpośrednio dostępne
    - Możemy je porównać wywołując funkcję ClusterProbability
  - CLUSTERING_METHOD
    - 3 (Scalable K-Means)
    - 4 (Non-scalable K-Means)

Microsoft Clustering – techniki segmentacji

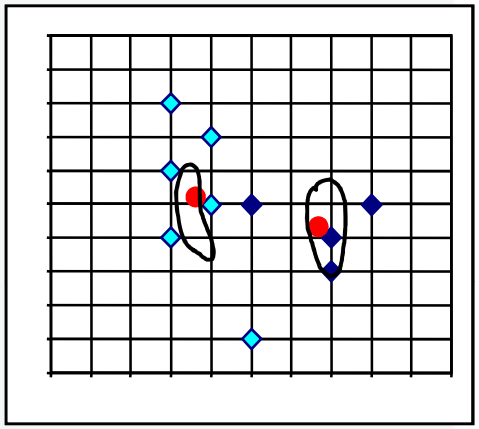
- K-Means



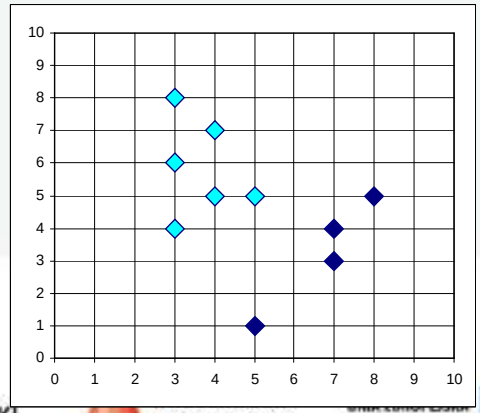
Przypisanie przypadków do klastrów



Przesunięcie klastrów

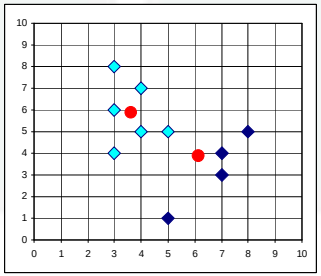


Ponowne przypisanie przypadków



Przesunięcie klastrów

Ponowne przypisanie przypadków



K=2
Wyznaczenie środków klastrów

Microsoft Clustering – techniki segmentacji

- EM (Expectation Maximization)
  - Technika modelowania statystycznego
  - Krok E – obliczenie prawdopodobieństwa przynależności przypadku do klastrów

$$w_h^j(x) = \frac{w_h^j \cdot f_h(x \mid \mu_h^j, \Sigma_h^j)}{\sum_i w_i^j \cdot f_i(x \mid \mu_i^j, \Sigma_i^j)}$$

  - Krok M – zmiana parametrów modelu

$$\mu_h^{j+1} = \frac{\sum_{x \in D} w_h^j(x) \cdot x}{\sum_{x \in D} w_h^j(x)} \quad \Sigma_h^{j+1} = \frac{\sum_{x \in D} w_h^j(x)\left(x - \mu_h^{j+1}\right)\left(x - \mu_h^{j+1}\right)^T}{\sum_{x \in D} w_h^j(x)} \quad h = 1, \ldots, k$$

  - Przepadek z reguły trafia do kilku klastrów
  - CLUSTERING_METHOD
    - 1 (Scalable EM)
    - 2 (Non-scalable EM)

Microsoft Clustering - podsumowanie

- Użycie PREDICT ONLY oznacza, że atrybut nie będzie używany dla klastrowania, ale zostanie dodany do klastrów po ich wyznaczeniu
- Przykładowe zastosowania:
  - Przygotowanie danych do dalszej eksploracji
  - Wykrywanie anomalii i oceny ich prawdopodobieństwa
    - W przypadku danych ciągłych oceniane jest prawdopodobieństwo ich dystrybucji
  - Segmentacja
  - Predykcja
    - Nieznany atrybut nie jest używany do klastrowania
    - Po przydzieleniu przypadku do klastra odczytywana jest jego struktura

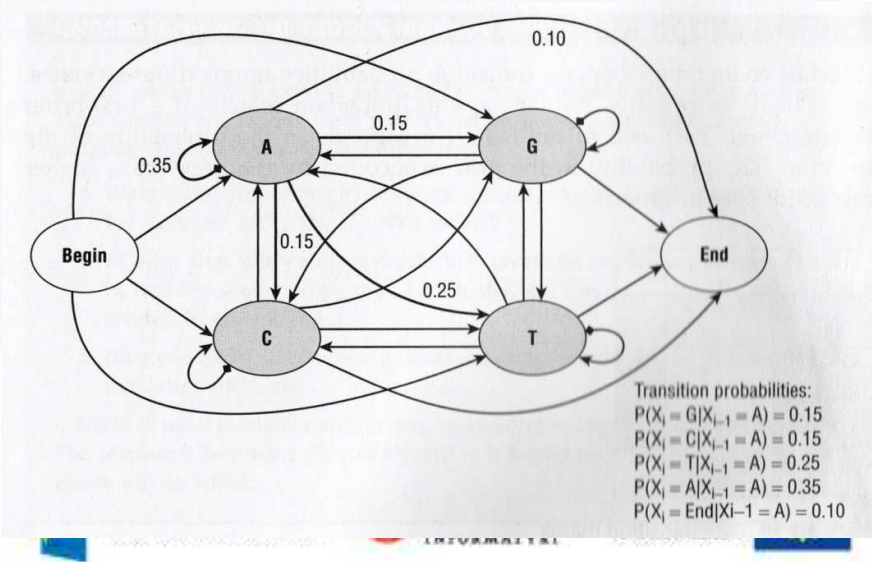
Microsoft Sequence Clustering

- Sekwencje - specjalny typ zagnieżdżonej tabeli:
  - Zawierającej kolumnę klucza sekwencji
  - I atrybuty sekwencji (np. adres URL odwiedzonej strony)

ID klienta	Wiek	Stan cywilny	Zakupy samochodów	
			ID sekw	Marka
1	35	ZŻ	1	Porch-A
			2	Bamborgini
			3	Kexus
2	20	W	1	Wagen
			2	Voovo
			3	Voovo
3	57	ZŻ	1	Voovo
			2	T-Yota

Microsoft Sequence Clustering

- Łańcuch Markowa – ciąg zdarzeń, w którym prawdopodobieństwo każdego zdarzenia zależy jedynie od wyniku poprzednich zdarzeń
  - W tym przypadku tylko od prawdopodobieństwa poprzedniego zdarzenia
- Macierz przejścia - rozkład prawdopodobieństw przejść między poszczególnymi stanami

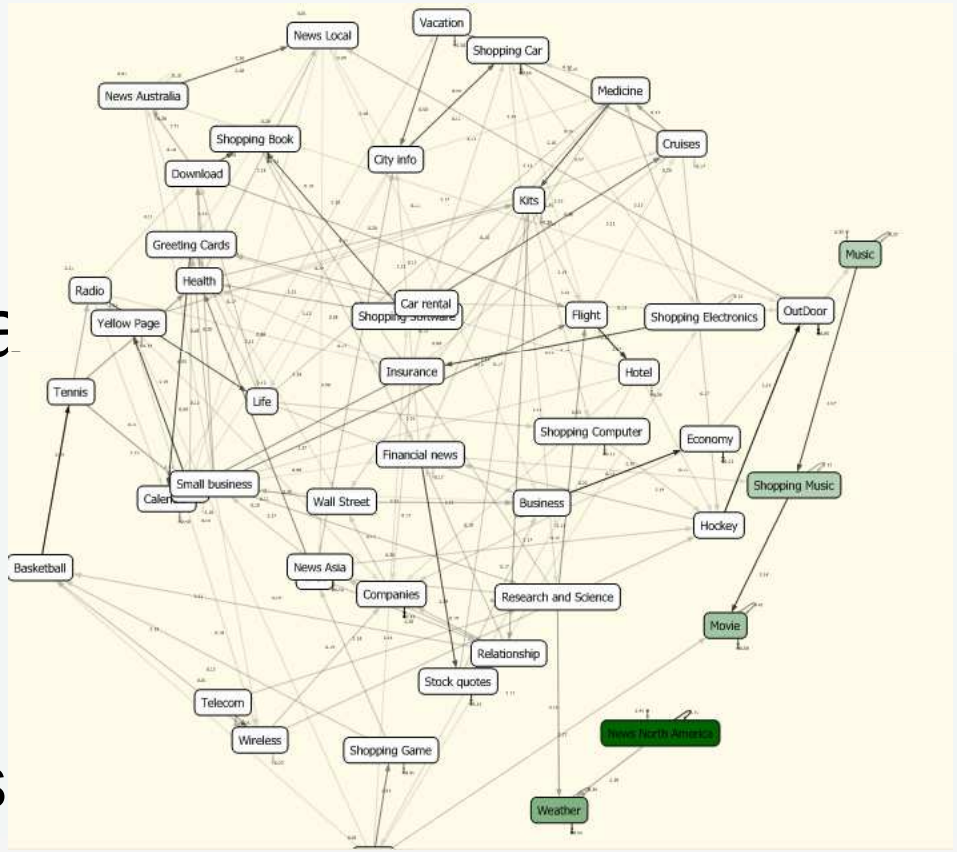


Microsoft Sequence Clustering

- Połączenie segmentacji z analizą sekwencyjną
  - Tworzone są klastry
  - Dla klastrów generowane są jawne łańcuchy Markowa
  - Każdy przypadek przypisywany jest z określonym prawdopodobieństwem do klastrów (krok E)
    - Uwzględniane są prawdopodobieństw przejść pomiędzy poszczególnymi stanami
    - $P(x|C) = P(xL|xL_{-1})\,P(xL_{-1}|xL_{-2})\,\ldots\,P(x_2|x_1)P(x_1)$
  - Następuje zmiana parametrów modelu (krok M)
  - Jeżeli klastrów jest mało a generują one różne łańcuch Markowa, następuje ich dekompozycja
    - Parametr CLUSTER_COUNT

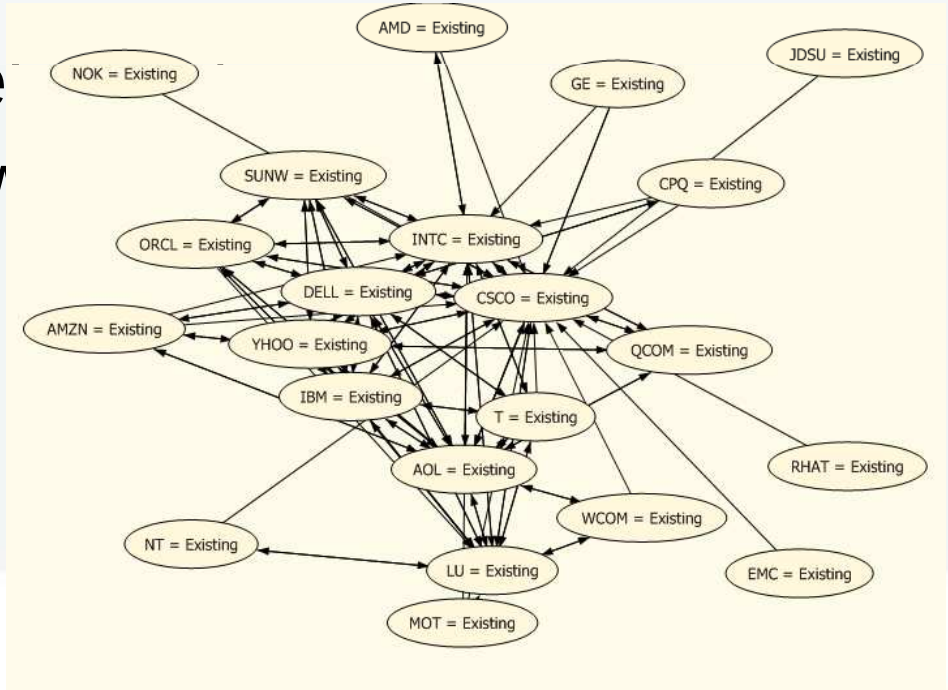
Microsoft Sequence Clustering - podsumowanie

- Mieszanka technologii klastrowania i sekwencjonowania
  - Grupowanie osób na podstawie ich profili w tym danych sekwencyjnych (prawdopodobieńs w zmiany stanów)



Microsoft Associaton Rules

- Klasyczny algorytm asocjacji apriori
  - Czyli zliczający stany atrybutów
- W pierwszej kolejności wyszukiwane są popularne zbiory
  - Najpierw zbiory 1-e
  - Potem 2-elementov
  - Itd.



Microsoft Associaton Rules

- Liczba transakcji w których występują szukane zbiory nazywa się wsparciem
  - Wsparcie ({A,B}) = Liczba transakcji (A,B)
- Wsparcie (popularność szukanych zbiorów) można ograniczyć:
  - MINIMUM_SUPPORT
    - Zbyt mała wartość parametru (poniżej 1 %) powoduje wykładniczy wzrost wymaganej pamięci i mocy obliczeniowej
  - MAXIMUM_SUPPORT
  - Progi bezwzględne lub procentowe

Microsoft Associaton Rules

- W drugiej kolejności generowane są reguły
  - Pewność określa minimalne prawdopodobieństwo wystąpienia reguły
    - Pewność A → B = Wsparcie (A,B) / Wsparcie (A)
    - MINIMUM_PROBABILITY
  - Ważność (podniesienie) określa niezależność elementów
    - Ważność {A,B} = Praw. (A,B) / Praw. (A) * Praw. (B)
      - Ważność (A → B) = log (Praw. (B/A)/ Praw. (B/ not A)
    - MINIMUM_IMPORTANCE
      - 0 oznacza, że elementy są niezależne
      - Wartości ujemne – że są negatywnie skorelowane
      - Wartości dodatnie – że elementy są pozytywnie skorelowane

Microsoft Associaton Rules

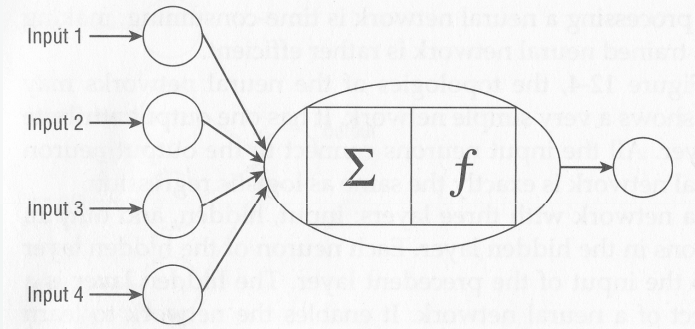
- Umożliwia predykcje
  - Jeżeli dla zbioru nie istnieje reguła, używane są reguły istniejące dla podzbiorów
  - Jeżeli dla podzbiorów również nie ma pewnych reguł, wskazywane są najpopularniejsze elementy

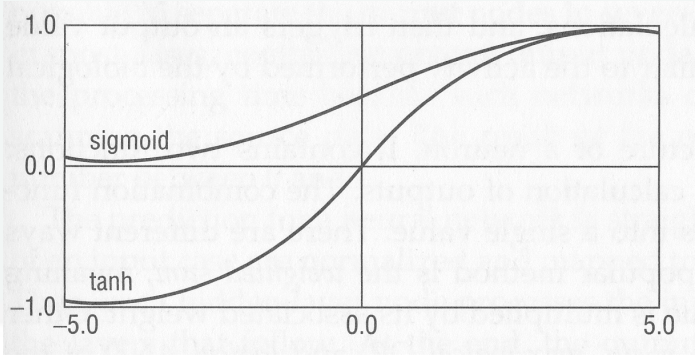
Microsoft Associaton Rules - podsumowanie

- Rozwiązuje jasno postawione zadania
- Łatwe przygotowanie danych
  - Wystarczy zadbać o reprezentatywność próbki
  - Wspiera zagnieżdżone przypadki
- Zwraca intuicyjne i łatwe do praktycznego zastosowania wyniki
- Łatwa ocena modelów
  - Uzyskanie fałszywych wyników jest prawie niemożliwe
  - Nieprzydatne wyniki są łatwe do wykrycia

Microsoft Logistic Regression

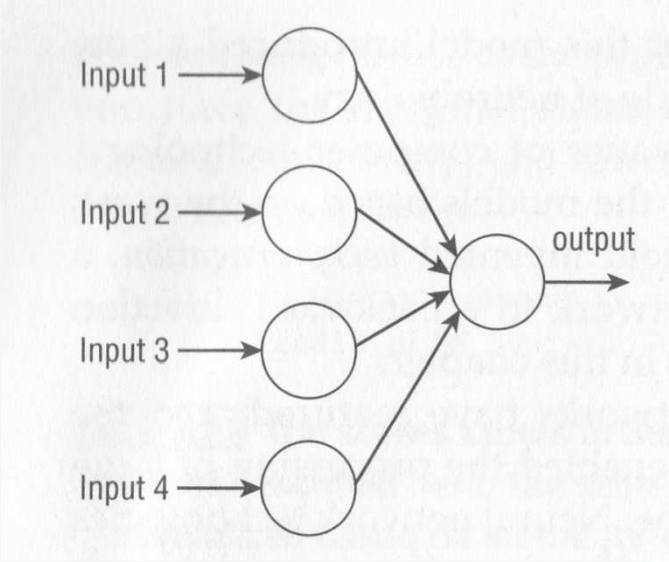
- Technika stochastyczna
  - Czyli „zgadująca” przybliżone wyniki
- Znajduje zależność pomiędzy atrybutami wejściowymi a przewidywanym
- Automatycznie normalizuje wartości atrybutów wejściowych





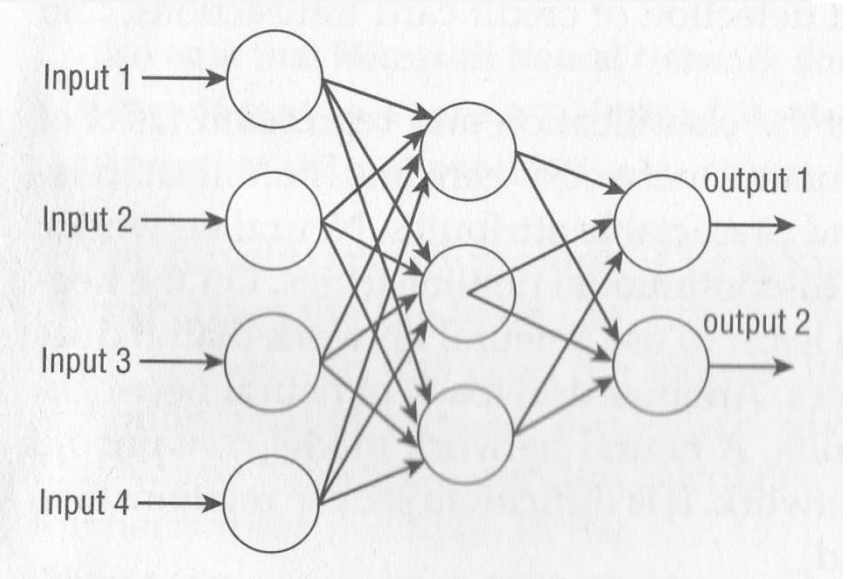
Microsoft Neural Network

- Sieć składa się z węzłów (neuronów) i łączy
  - Każdy węzeł jest jednostką przetwarzania danych
    - Węzły mają wiele wejść i jedno wyjście
    - Sygnał na wyjściu jest sumą ważoną sygnałów wejściowych
  - Każdy atrybut wejściowy jest reprezentowany przez osobny węzeł



Microsoft Neural Network

- Każdy węzeł wyjściowy reprezentuje jeden przewidywany atrybut
  - W sieciach bez ukrytej warstwy węzły wejściowe są bezpośrednio powiązane z wyjściowymi
  - W sieciach z ukrytymi warstwami parametry wejściowe są łączone ze sobą
  - Używany jest tylko jeden ukryty poziom



Microsoft Neural Network

- W pierwszej kolejności określana jest struktura sieci
  - HIDDEN_NODE_RATIO
- Następnie szukane są optymalne wagi łączy
  - Otrzymane wyniki są porównywane ze znanymi danymi
  - Błędy są eliminowane przez zmianę wagi łączy
  - Proces jest powtarzany aż do:
    - Osiągnięcia wystarczającej dokładności
    - Przekroczeniu liczby iteracji
    - Przekroczeniu skali zmiany wag łączy

Microsoft Logistic Regression - podsumowanie

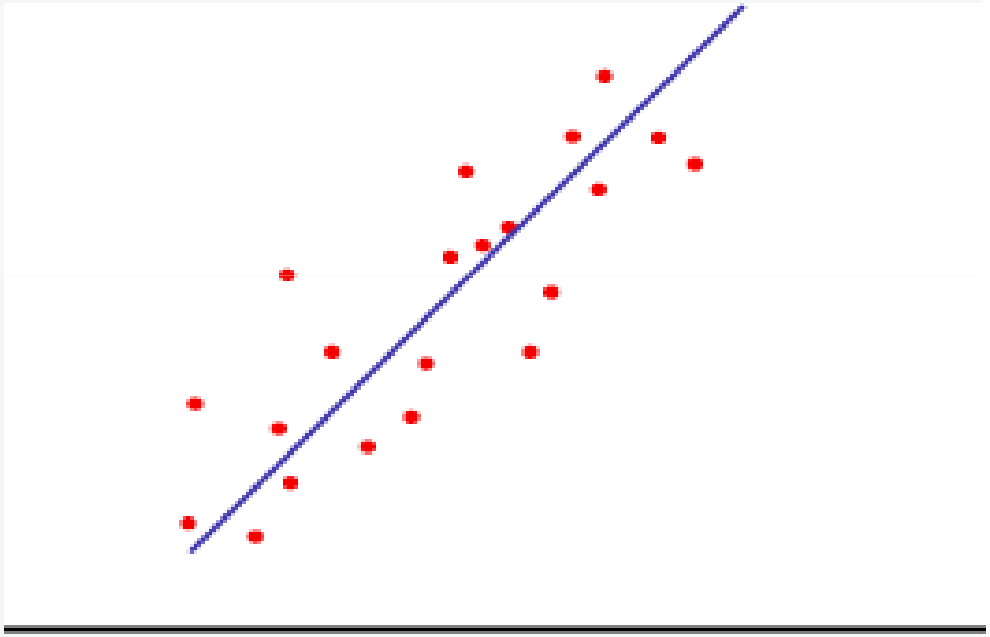
- Sieci neuronowe bez ukrytej warstwy
- Znacznie szybszy od Microsoft Neural Network
  - Ale z reguły równie dokładny
- Kolejny klasyfikator liniowy
  - Jeśli przewidywany parametr zależy od kombinacji wartości parametrów wejściowych, zależność nie zostanie znaleziona

Microsoft Neural Network - podsumowanie

- Niezastąpione w znajdowaniu skomplikowanych relacji pomiędzy atrybutami
  - Również w dużych zbiorach danych
- Kosztowny trening
- Trudne do interpretacji wyniki

Microsoft Linear Regression

- Modyfikacja algorytmu drzew decyzyjnych służącą do regresji wartości ciągłych
- Używany do predykcji
- Znajduje najlepszą linie prostą przybiegającą przez serie punktów

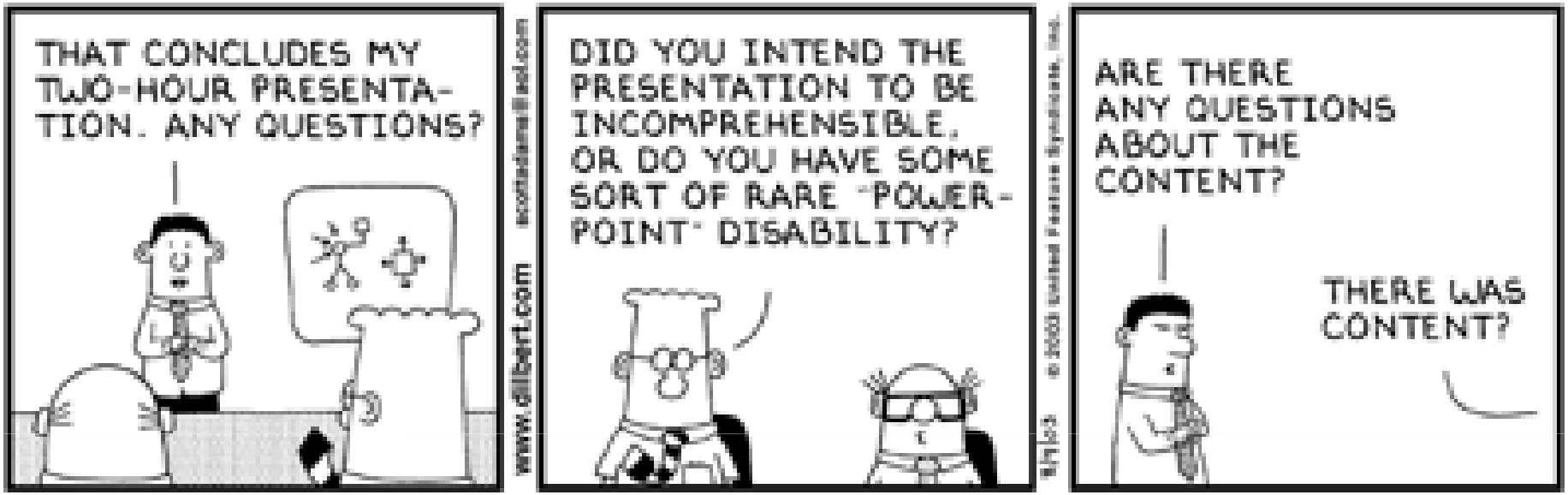


Microsoft Linear Regression - podsumowanie

- Algorytm drzew decyzyjnych zakłada (na podstawie danych treningowych) określoną liczbę sytuacji powodujących podział drzew decyzyjnych. Natomiast algorytm regresji liniowej nie jest ograniczony przez dane treningowe i może służyć do przewidywania wyników wykraczających poza zakres danych historycznych.

Algorytmy eksploracji danych

Algorytm	Opis
Drzewa decyzyjne	Określa szanse wyniku na podstawie wartości w zestawie treningowym
Reguły asocjacyjne	Określa relacje między przypadkami
Klastrowanie	Klasyfikuje przypadki na odrębne grupy na podst. zbiorów atrybutów
Naiwny klasy-fikator Bayesa	Wyraźnie przedstawia różnice w konkretnej zmiennej dla różnych elementów danych
Klastrowanie sekwencyjne	Grupuje lub klastruje dane na podstawie sekwencji poprzednich zdarzeń
Szeregi czasowe	Analizuje i prognozuje dane czasowe łącząc możliwości rozwiązania ARTXP (opracowanego przez zespół Microsoft Research) do krótkoterminowych przewidywań z metodą ARIMA (w SQL 2008) w celu osiągnięcia precyzji w dłuższej perspektywie.
Sieci neuronowe	Szuka nieznanych nieintuicyjnych relacji w danych
Regresja liniowa	Określa relację między kolumnami w celu przewidywania wyniku
Regresja logistyczna	Określa relację między kolumnami w celu oceny prawdopodobieństwa, że kolumna będzie zawierać konkretny stan



Marcin@wss.pl