

Podstawy sztucznej inteligencji

wykład 7

Eksploracja danych

Joanna Kołodziejczyk

09 stycznia 2013

Plan wykładu

- 1 Co to jest eksploracja danych?
- 2 Jak działa eksploracja danych?
- 3 Metody w eksploracji danych

Definicja

Eksploracja danych ED (Data mining)

Metody wydobywania ukrytych informacji z dużych baz danych.

Cel

Do prognozowania przyszłych trendów i zachowań, które pozwolą przedsiębiorstwom na podejmowanie opartych na wiedzy decyzji.

Zalety

- Zautomatyzowana prospektywna analiza danych wykracza poza zwykłe narzędzia wspomagania decyzji.
- ED udziela odpowiedzi na pytania, które nie znajdowały odpowiedzi ze względu na złożoność obliczeniową.
- Poszukują w bazach danych ukrytych wzorców, informacji, które ekspert może pominąć, gdyż znajdują się poza jego oczekiwaniami.

Technologie pozwalające na rzeczywiste wykorzystanie ED

Zasoby zapewniające wykorzystanie ED:

- olbrzymie i prawie wszechobecne zbiory danych
- zwiększająca się moc obliczeniowa komputerów
- algorytmy eksploracji danych.

Technologie eksploracji danych wywodzą się z obszarów badań:

- statystyka
- sztuczna inteligencja
- maszynowe uczenie się.

Zakres zastosowania eksploracji danych

Automatyczne przewidywanie trendów i zachowań

Automatyzuje się proces wyszukiwania informacji i można szybko udzielać odpowiedzi na pytania dotyczące danych.

Przykłady:

- Ukierunkowany marketing: wykorzystanie np. danych z przeszłych korespondencji promocyjnych do określenia klientów maksymalizujących szansę ponownych inwestycji.
- Prognozowanie upadłości: identyfikacja segmentów biznesu, które mogą reagować podobnie na pewną sekwencję zdarzeń.

Zakres zastosowania eksploracji danych

Automatyczne wykrywanie nieznanych wcześniej wzorców

Narzędzia eksplorują bazy danych i identyfikują ukryte wzorce.
Przykłady odkrywania wzorców

- Analiza danych o sprzedaży detalicznej do identyfikacji pozornie niepowiązanych produktów, które często są nabywane razem.
- Wykrywanie wzorca fałszywych transakcji z użyciem kart kredytowych.
- Identyfikacja anomalii w danych.

Plan wykładu

- 1 Co to jest eksploracja danych?
- 2 Jak działa eksploracja danych?
- 3 Metody w eksploracji danych

Zadania wykonywane w ramach eksploracji danych

- Klasyfikacja
- Grupowanie
- Asocjacje
- Wzory sekwencyjne

Klasyfikacja

Definicja

Dane układa się w ustalonych grupach (klasach). Dane składają się z atrybutów obiektów i przypisanych im etykiet klas. Jeżeli pojawi się nowy obiekt o znanych atrybutach a nieznaney klasie zostaje on dopasowany do jednej z klas. Modelowanie zależności wejście-wyjście nazywane jest uczeniem nadzorowanym.

Przykłady:

- Filtrowanie spamu (niechcianyach wiadomości).
- Zwiększenie konkurencyjności restauracji poprzez ustalanie specjalności dnia na podstawie zamówień klientów. Możliwe jest określenie kiedy najczęściej klienci odwiedzają lokal i co zazwyczaj zamawiają.

Grupowanie

Definicja

Dane są grupowane według logicznych powiązań. Dane składają się z atrybutów obiektów i nie posiadamy przypisanych im etykiet klas. Obiekty łączą się w grupy według pewnej przyjętej relacji. Nowy obiekt zostaje przypisany do pewnej grupy przez podobieństwo do grupy. Modelowanie grup nazywa się uczeniem nienadzorowanym.

Przykłady:

- Bardzo często wykorzystywane przy pracy z danymi ankietowymi i testach. Wykorzystuje się grupowanie, by całą populację klientów podzielić na mniejsze grupy tak, by zidentyfikować działy rynku i lepiej zrozumieć zależności pomiędzy różnymi grupami klientów, potrzebami rynku. Stosuje się do określenia polityki marketingowej.
- Grupowanie produktów w sklepach internetowych.

Asocjacje

Definicja

Dane służą do identyfikacji związków pomiędzy atrybutami. W dużych bazach danych poszukuje się reguł, które określają silne (według przyjętego kryterium) powiązania pomiędzy cechami obiektu.

Przykłady:

- Reguła asocjacyjna: piwo-> pieluchy, cebula, ziemniaki -> mięso. Wykorzystywane do rozmieszczanie produktów i akcji promocyjnych.
- Wykrywanie włamań komputerowych. Wykrywanie reguł, które łączy się z atakiem. Wykonuje się to poprzez analizy tysięcy linii logów i poszukiwaniu anomalii.

Wzory sekwencyjne

Definicja

Dane wykorzystuje się do przewidywania zachowań i trendów. Dane pojawiają się sekwencyjnie i przechowywane są w sposób wskazujący na kolejność ich pojawiania się. Poszukuje się w nich statystycznie istotnych wzorców.

Przykłady:

- Sprzedawcy sprzętu mogą przewidzieć prawdopodobieństwo nabycia ubezpieczenia w pewnym czasie po zakupie komórki na podstawie zakupu konkretnego typu telefonu komórkowego. W pewnym sensie jest to wykrycie reguły, w której ważne jest następstwo czasowej
- Wykrywanie brakującego fragmentu DNA lub ciągu znaków.

Techniki eksploracji danych

- **sztuczne sieci neuronowe**
- **drzewa decyzyjne**: struktura drzewiasta, które zawiera zestawy decyzji. Decyzje te generują zasad klasyfikacji zbioru danych. Metody wykorzystujące drzewa decyzyjne to drzewa klasyfikacyjne i regresyjne.
- **algorytmy genetyczne** do wykrywania reguł w danych
- **metoda najbliższego sąsiedztwa**: technika, która grupuje rekordy w zbiorze danych łącząc ze sobą k rekordów najbliższych (podobnych do niego) dla pewnego wybranego rekordu.
- **indukcja reguł**: wydobywanie reguł (jeśli-to) w oparciu o istotność statystyczną.

Modelowanie

Modelowanie

jest to tworzenie modelu dopasowanego do pewnej sytuacji, w której znane jest zachowanie/odpowiedź i zastosowanie go do innej sytuacji, gdy odpowiedź nie jest znana.

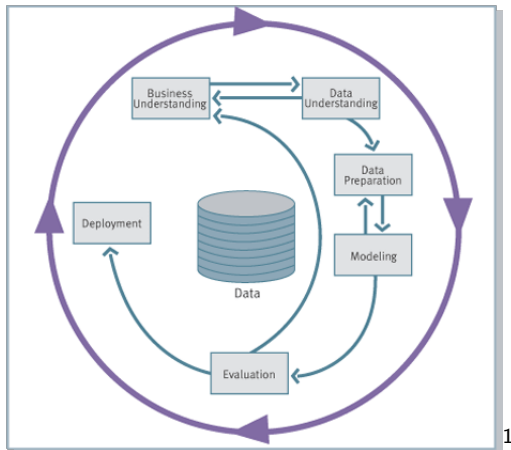
Przykład firmy telekomunikacyjnej

Z danych historycznych o usługobiorcach zostanie zbudowany model, który określi potencjalnych klientów rozmów międzynarodowych. Modelowanie odgaduje zależności istniejące w bazie danych i tak możliwy model to:
98% klientów, którzy zarabiają więcej niż 60.000 rocznie wydaje więcej niż 80/miesiąc na rozmowy międzynarodowe.

Przykłady zastosowań

- Firma transportowa może określać najlepsze perspektywy dla swojej działalności na podstawie eksploracji danych. Analizując doświadczenia z klientami można wyznaczyć segmenty działalności (wyznaczyć atrybuty) o największym wpływie na przyszłą działalność. Można takie wyniki uogólnić na cały region.
- Firmy prowadzące sprzedaż mogą próbować zwiększać wskaźniki sprzedaży wykorzystując eksplorację danych. Dane z paneli konsumenckich, dostaw, aktywności konkurencji pozwalają zrozumieć trendy w zmianach marki i sklepów. Producent na tej podstawie może planować kampanię reklamową i najlepsze sposoby dotarcia do klienta.

Proces eksploracji danych



¹źródło: <http://www.crisp-dm.org/Process/index.htm>

Plan wykładu

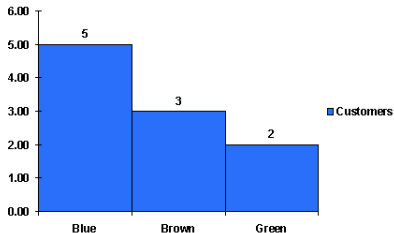
- 1 Co to jest eksploracja danych?
- 2 Jak działa eksploracja danych?
- 3 **Metody w eksploracji danych**
 - Statystyka w eksploracji danych
 - Najbliższe sąsiedztwo
 - Klasteryzacja
 - Drzewa decyzyjne

Statystyka

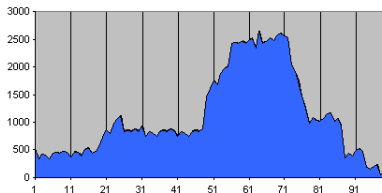
Używając narzędzi ze statystyki można udzielać odpowiedzi na pytania:

- Jakie wzorce są ukryte w bazie danych?
- Jaka jest szansa, że nastąpi pewne zdarzenie?
- Jakie wzorce są istotne?
- Co wynika z „podsumowania” (np. średnia) danych? Zyskuje się pewne wyobrażenie o tym, co jest zawarte w bazie danych.

Histogramy



kolor oczu



wiek

Użyteczne miary

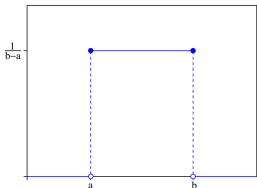
- **Max** - maksymalna wartość z danych.
- **Min** - minimalna wartość z danych.
- **Średnia** - średnia wartość w próbie.
- **Mediana** - wartość w bazie, powyżej i poniżej której znajduje się jednakowa liczba rekordów (dzieli bazę na połówki o równej liczbie rekordów).
- **Dominanta** - wartość najczęściej występująca (o największym prawdopodobieństwie wystąpienia).
- **Wariancja** - miara zmienności, tego, jak rozkładają się wartości od wartości średniej.

Rozkłady

Czasami zamiast histogramu chce się opisać rozkład danych równaniem. W klasycznej statystyce zakłada się, że istnieje pewien „prawdziwy”, podstawowy kształt rozkładu, który powstaje wtedy, gdy zostaną zebrane wszystkie możliwe dane.

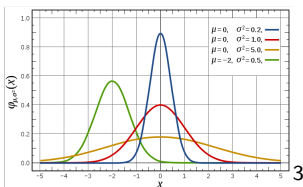
Zadaniem statystyka jest określenie prawdopodobnego rozkładu z ograniczonej liczby danych .

Wiele rozkładów opisanych jest tylko przez średnią i wariancję.



jednostajny

³źródło: wikipedia

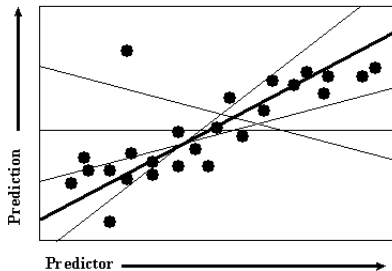


normalny

Regresja liniowa

Podstawowa zasada regresji jest taka, że z mapy wartości jest tworzony taki model, by uzyskać najniższy błąd (zazwyczaj średniokwadratowy).

$$\text{Prediction} = a + b \cdot \text{Predictor}$$



4

⁴źródło: <http://www.thearling.com/text/dmtechniques/dmtechniques.htm>



Bardziej złożone modele niż liniowe

Złożoność modelu może wynikać z:

- zwiększenia liczby wejść (predictors) (zwiększenie wymiarowości)

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$$

- regresja nieliniowa — zastosowania przekształcenie dla wejścia (podnoszenie do potęgi)

$$Y = a + b_1X_1 + b_2X_1^2$$

- wymnażania przez siebie wejść
- modyfikacji by odpowiedź modelu była binarna (regresja logistyczna)

Grupowanie

Klasyfikacja metodą najbliższego sąsiada

Zasada polega na tym, że jeżeli chcę wiedzieć jaka jest odpowiedź (prognozowane wyjście) na sygnał wejściowy, to patrzę na najbliższe sąsiednie rekordy o podobnych wejściach z danych historycznych i używam taką samą klasę.

Przykład grupowania

Grupowanie odzieży do prania, czyszczenia. Grupuje się je, gdyż mają podobną charakterystykę.

Grupowanie przez najbliższe sąsiedztwo

Przykład: prawdopodobnie większość Twoich sąsiadów (sąsiedztwo geograficzne) ma podobny przychód. Metoda ta jest intuitywna a jednocześnie łatwa do zautomatyzowania.

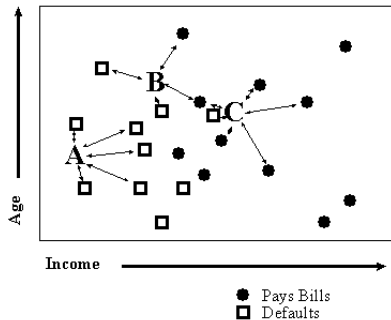
Metoda najbliższego sąsiada w predykcji

U podstaw koncepcji klastrów (grup) leży to, że dany obiekt (czy to samochody, żywność lub klient) może być bliżej do innego obiektu, niż jakiś inny trzeci obiekt. Większość ludzi ma wrodzone poczucie porządkowania różnych przedmiotów i zgodzi się, że jabłku bliżej do pomarańczy niż do pomidora. To poczucie pozwala nam budować klastry - zarówno w bazach danych, jak również w codziennym życiu. Definicja bliskości pozwala również dokonać prognozy.

Sąsiedztwo do predykcji

Obiekty leżące blisko siebie powinny mieć taką samą wartość predycyjną. Wystarczy zatem znać wartość wyjściową dla jednego obiektu.

K-najbliższych sąsiadów do klasyfikacji



5

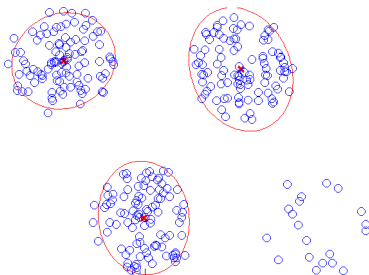
Zaufanie do predykcji

Tym większa wiarygodność im bliższe sąsiedztwo lub jednorodność K-sąsiadów.

⁵źródło: <http://www.thearling.com/text/dmtechniques/dmtechniques.htm>

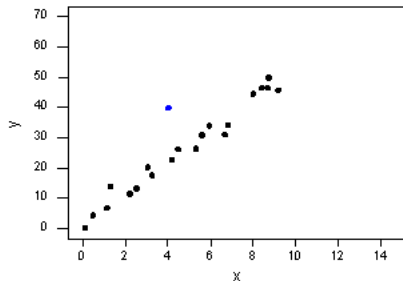
Grupowanie nienadzorowane (bez znanej odpowiedzi)

W danych nie ma predykcji. Grupowanie polega na obserwacji rozkładu danych w przestrzeni wejść i nadawanie tej samej etykiety dla blisko sąsiadujących rekordów separowalnych od innych.



Grupowanie pozwala wychwycić odstające rekordy (outliers)

Dzięki grupowaniu można łatwo zidentyfikować odstające rekordy i wskazać przyczynę tego stanu rzeczy.



Różne wyniki grupowania

Według przychodu

ID	Name	Prediction	Age	Balance	Income	Eyes	Gender
3	Betty	No	47	\$16,543	High	Brown	F
5	Carla	Yes	21	\$2,300	High	Blue	F
6	Carl	No	27	\$5,400	High	Brown	M
8	Don	Yes	46	\$0	High	Blue	M
1	Amy	No	62	\$0	Medium	Brown	F
2	Al	No	53	\$1,800	Medium	Green	M
4	Bob	Yes	32	\$45	Medium	Green	M
7	Donna	Yes	50	\$165	Low	Blue	F
9	Edna	Yes	27	\$500	Low	Blue	F
10	Ed	No	68	\$1,200	Low	Blue	M

6

Różne wyniki grupowania

Według wieku i koloru oczu

ID	Name	Prediction	Age	Balance	Income	Eyes	Gender
5	Carla	Yes	21	\$2,300	High	Blue	F
9	Edna	Yes	27	\$500	Low	Blue	F
6	Carl	No	27	\$5,400	High	Brown	M
4	Bob	Yes	32	\$45	Medium	Green	M
8	Don	Yes	46	\$0	High	Blue	M
7	Donna	Yes	50	\$165	Low	Blue	F
10	Ed	No	68	\$1,200	Low	Blue	M
3	Betty	No	47	\$16,543	High	Brown	F
2	Al	No	53	\$1,800	Medium	Green	M
1	Amy	No	62	\$0	Medium	Brown	F

7

Problemy z grupowaniem

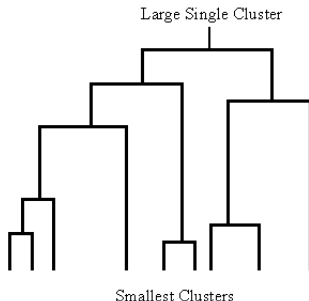
- Który rekord do którego klastra? Algorytm podziału na klastry powinien mieć określone zasady, jaka cecha ma większy priorytet i jaki atrybut jest ważniejszy.
- Kompromis liczności klastrów i jednorodności. Chcąc uzyskać najbardziej jednorodne klastry będziemy mieli tendencję do zwiększania liczby klastrów (aż do liczby rekordów). Natomiast chcąc uzyskać generalizację trzeba dla danego problemu próbować budować jak najmniej klastrów.

Porównanie klasyfikacji i grupowania

Klasyfikacja	Grupowanie
Służy do prognozowania, jak również konsolidacji.	Używana głównie do konsolidowania danych (widok z góry na przestrzeń wejść) i zapisu do grup.
Przestrzeń jest zdefiniowana przez problem (uczenie nadzorowane).	Przestrzeń jest zdefiniowana jako domyślna przestrzeń n-wymiarowa lub zdefiniowana przez użytkownika, lub jest predefiniowaną przestrzenią dostarczoną przez wcześniejsze doświadczenia (uczenie bez nadzoru).
Używa metod metrycznych do określenia bliskości rekordów.	Może używać inne niemetryczne miary.

Grupowanie hierarchiczne

Metody hierarchiczne tworzą podziały na różne liczności klastrów. Istnieje możliwość decydowania o wygodnym doborze liczby klastrów.



8

Metody grupowania hierarchiczne

- Poprzez łączenie (aglomerative) — techniki grupowania zaczynające od liczby klastrów równej liczbie rekordów. Klastry, które znajdują się najbliżej siebie są łączone ze sobą tworząc drugi co do wielkości klastr. To połączenie jest kontynuowane aż do utworzenia jednego klastra zawierających wszystkie rekordy, znajdującego się na szczycie hierarchii.
- Poprzez podziały (divisive) — techniki grupowania działające w odwrotnym kierunku niż powyższa technika. Zaczynają gdy wszystkie rekordy są zgrupowane w jeden klastr, a następnie dokonują podziału na mniej liczne grupy.

Grupowanie niehierarchiczne

Metody są zdecydowanie szybsze od hierarchicznych, ale wymagają od użytkownika podania:

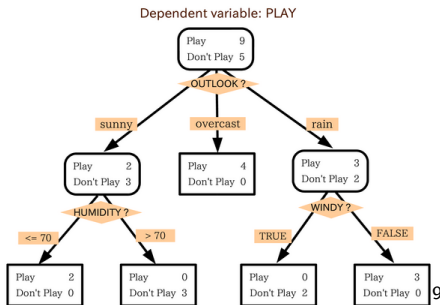
- pożądanej liczby klastrów lub
- minimalnej wymaganej bliskość dwóch rekordów w jednym klastrze.

Często wykonują się iteracyjnie startując z inną początkową konfiguracją rekordów, która wpływa na ostateczny podział oraz dokonują w pętli poprawek na granicach klastrów.

Drzewo decyzyjne

Drzewo decyzyjne

Jest to model predykcyjny w formie drzewa. Każda gałąź jest odpowiedzią na pytanie o klasyfikację, o której decyzja zawarta jest w liściu.



⁹źródło:

<http://gautam.lis.illinois.edu/monkmiddleware/public/analytics/decisiontree.html>

Cechy drzewa decyzyjnego

- Dzieli się dane w każdym punkcie podziału bez utraty danych (łączna liczba pozycji w węźle rodzicu jest równa sumie zapisów zawartych w jej potomkach).
- Łatwo jest zrozumieć jak powstaje model (w przeciwieństwie do sieci neuronowych czy klasycznej statystyki).
- Model zgodny z intuicją.
- Drzewo decyzyjne może być postrzegane jako tworzenie segmentów (klientów, produktów, regionów sprzedaży). Segmenty są tworzone poprzez podobieństwo rekordów wynikające z ich przynależności do zmiennej predykcyjnej.

Zastosowania drzew decyzyjnych

- Algorytmy budują pełne drzewo dla hipotezy. Odtwarzają sposób analizy problemu przez specjalistę. Dla dużych rzeczywistych problemów, mogą być bardzo złożone.
- Służą do eksploracji danych. Dokonuje się ona przez patrzenie na zmienną decyzyjną i zmienną podziału w drzewie. Np. Jeśli klient ma umowę $< 1,1$ roku i kanał sprzedaży = telesprzedaż THEN możliwość rezygnacji wynosi 65%.
- Do wstępnej obróbki danych przed predykcją np. do wyznaczania istotnych wejść do sieci neuronowej.
- Do predykcji.

Algorytmy

- ID3 — rozdziela atrybuty na podstawie miar informacyjnych (entropii).
- C4.5 — udoskonalenie ID3: zmniejszenie liczby obliczeń, możliwość użycia zmiennych ciągłych, praca z atrybutami z brakującymi wartościami.
- CART (Classification And Regression Tree) — udoskonalenie C4.5. Stosuje „node impurity” do wskazania atrybutu podziału w drzewie.

Szczegóły: <http://courtdecisionsandrulings.com/an-integrated-study-on-decision-tree-induction-algorithm.html>