Spis treści

Eksploracja danych – Notatki na egzamin

Pojęcia wstępne

Eksploracja danych

Eksploracja danych jest procesem odkrywania ważnych nowych współzależności wzorców, trendów dzięki przeszukiwaniu dużych ilości danych przechowywanych w bazach za pomocą zarówno technik rozpoznawania wzorców jak i metod statystycznych i matematycznych. Zastosowanie eksploracji danych znajdziemy w ocenie ryzyka, migracji klientów, wykrywaniu nadużyć, analizie koszyka zakupowego, segmentacji klientów i web miningu.

Klasyfikacja

Przypisanie wartości pewnej jakościowej zmiennej celu na podstawie zmiennych opisujących. Algorytm klasyfikacji działa tak, że najpierw uczy się na pewnym podzbiorze zwanym uczącym jakim kombinacjom zmiennych opisujących odpowiadają konkretne wartości zmiennej celu, a następnie przewiduje wartość zmiennej celu dla rekordów, dla których nie jest ona znana.
Metody Klasyfikacyjne:

1. Algorytm K-NN
2. Drzewa klasyfikacyjne
3. Lasy losowe
4. Sieci neuronowe

Przykładami w przykładzie klasyfikacji może być określenie czy dana transakcja karta kredytowa jest oszustwem czy nie, bądź identyfikacja zagrożeń terrorystycznych na podstawie zachowania i finansów osób.

Szacowanie (regresja)

Budowanie modelu szacujanego wartości ilościowej zmiennej celu na podstawie innych dostępnych zmiennych. Następnie dla nowych obserwacji szacuje się wartości zmiennej celu, opierając się na wartościach zmiennych opisujących.
Metody Szacujące:

1. Przedziały ufności
2. Regresja
3. Sieci Neuronowe

Przykładem jest np. przewidzenie ceny domu na podstawie jego lokalizacji, powierzchni, roku budowy itp.

Przewidywanie

Jest to zagadnienie podobne do szacowania i klasyfikacji, z tym, że wynik dotyczy przyszłości.
Użyteczne metody to:

1. Regresja
2. sieci neuronowe
3. drzewa decyzyjne
4. metoda K-NN

Przykładami może być przewidywanie procentowego wzrostu liczby ofiar śmiertelnych w przyszłym roku po zwiększeniu dozwolonej prędkości poruszania się po drodze lub wzrostu zapotrzebowania na opony zimowe.

Grupowanie

Z angielskiego clustering oznacza grupowanie rekordow w klasy obiektów na podobnych do siebie, a niepodobnych do innych grup. Różni się tym od klasyfikacji, że nie ma zmiennej celu. Stosowane metody:

1. Grupowanie hierarchiczne
2. K-średnich
3. Sieci Kohonena

Przykładami może być namierzenie grupy potencjalnych klientów pewnego produktu z niszy rynkowej wyprodukowanego przez firmę z małym budżetem reklamowym lub profil demograficzny różnych regionów kraju.

Odkrywanie reguł

Odkrywanie reguł polega na poszukiwaniu nieodkrytych reguł do ilościowego określenia relacji pomiędzy dwoma lub więcej atrybutami. Metody stosowane do tego to:

1. Algorytm a priori

Przykład: z 1000 klientów robiących zakupy w czwartek w nocy 200 kupiło pieluchy, a 50 z nich również piwo. Reguła: Jeżeli kupuje pieluchy, to kupuje piwo ma pokrycie 20%, wsparcie 5% i ufność 25%

Normalizacja min-max

$$X^* = \frac{X - \min X}{\max X - \min X}$$

Standaryzacja Z-score

$$X^* = \frac{X - \text{mean(X)}}{s_X}$$

Gdzie Sx to odchylenie standardowe

Rodzaje zmiennych

1. **Nominalna** – jest to rodzaj skali pomiarowej, zmienne są na skali nominalnej, gdy przyjmują wartości / etykiety, dla których nie istnieje wynikające z natury danego zjawiska uporządkowanie. Nawet jeśli wartości zmiennej nominalnej wyrażone są liczbowo, to liczby te są jedynie umownymi identyfikatorami, więc nie można na nich wykonywać działań arytmetycznych ani ich porównywać.
   Przykład: Powiat zamieszkania, płeć
2. **Porządkowa** – Rodzaj skali pomiarowej. Zmienne są na skali porządkowej, gdy przyjmują wartości, dla których dane jest uporządkowanie/kolejność, jednak nie da się w sensowny sposób określić ani różnicy ani ilorazu między dwoma wartościami.
   Przykład: wykształcenie, kolejność zawodników na podium
3. **Ilościowa** – Jej wartości reprezentują uporządkowane kategorie ze znaczącą metryką, która umożliwia porównywanie odległości między wartościami.
   Przykład: wiek w latach lub przychód w tysiącach złotych.

Miary jakości:

rzeczywista \ przewidywana	0	1
0	#TN	#FP
1	#FN	#TP

Zmienna *X*		Wartości przewidywane	
		0	1
Wartości obserwowane	0	90	90
	1	80	240

1. **Trafność** – odsetek poprawnych klasyfikacji. Jest to suma tego co na przekątnej dzielona na wszystkie obserwacje:

$$\frac{\#TN + \#TP}{N} = \frac{90 + 240}{90 + 90 + 80 + 240} = 0.66$$

2. **Całkowity współczynnik błędu** – Odsetek błędnych klasyfikacji. 1- Trafność lub suma tego co poza przekątną dzielona przez wszystkie obserwacje:

$$\frac{\#FP + \#FN}{N} = \frac{90 + 80}{90 + 90 + 80 + 240} = 0.34$$

3. **Czułość** – odsetek poprawnie sklasyfikowanych przypadków pozytywnych, liczony jako:

$$\frac{\#TP}{\#FN \ + \ \#TP} = \frac{240}{80 + 240} = 0.75$$

4. **Specyficzność / Swoistość** – odsetek poprawnie sklasyfikowanych przypadków negatywnych.

$$\frac{\#TN}{\#FP + \#TN} = \frac{90}{90 + 90} = 0.50$$

5. Średni błąd bezwzględny – Mean Absolute Error – MAE

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}$$

6. Błąd średniokwadratowy – Mean-Squared Error – MSE

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}$$

Algorytm K-NN

Algorytm k najbliższych sąsiadów służy przede wszystkim do klasyfikacji ale może być tez wykorzystany do problemu szacowania. Jest przykładem uczenia leniwego opartego na analizie przypadku tzn. algorytm zapamiętuje cały zbiór uczący i klasyfikacja nowych elementów jest dokonywana poprzez porównanie z najbardziej podobnymi rekordami zbioru uczącego.
Każdy rekord zawiera wartości p predyktorów i dlatego może być traktowany jako punkt w przestrzeni p-wymiarowej. Badamy podobieństwo rekordów, licząc odległość pomiędzy nimi. Im mniejsza odległość tym bardziej podobne są rekordy.

Metryka Euklidesowa: $d(z,y) = \sqrt{\sum_{i=1}^{p}(z_i - y_i)^2}$

Metryka Miejska: $d(z,y) = \sum_{i=1}^{p}|z_i - y_i|$

Drzewa Decyzyjne CART

Drzewo decyzyjne, inaczej klasyfikacyjne to drzewo skierowane mające jedyny dający się wyróżnić wierzchołek nazywany korzeniem, będący wierzchołkiem początkowym drzewa. Wierzchołki w tych drzewach nazywamy węzłami a krawędzie gałęziami. Drzewa cart są ściśle binarne. Jeśli warunek jest spełniony kierujemy się na lewo, w przeciwnym przypadku w prawo.
Model ten stosowany jest do przewidywania zmiennej celu dla nowych obserwacji. Wypuszczana jest ona z korzenia drzewa a następnie przesuwa się po gałęziach w lewo lub prawo w zależności od tego czy spełnia określony w danym węźle warunek czy nie. Finalnie trafia do jednego z liści.

PRAWDA:

1. Węzeł początkowy drzewa jest nazywany korzeniem.
2. Drzewa CRT mogą być stosowane dla ciągłej zmiennej celu.
3. Zmienne używane do budowy drzewa mogą być jakościowe.
4. Ograniczenie głębokości drzewa jest jedną z metod zapobiegania przeuczeniu.

FAŁSZ:

1. Algorytm w każdym węźle dokonuje podziału, dla którego miara nieczystości jest największa.
2. Drzewo CRT nie zadziała w przypadku wystąpienia braku danych dla zmiennych użytych w węzłach do podziału.
3. Algorytm w każdym węźle dokonuje podziału, dla którego wartość kryterium Ginnego jest najmniejsza
4. Końcowe węzły drzewa nazywamy rodzicami.

MLP

Główną zaletą MLP jest to że mogą być używane do predykcji, estymacji i klasyfikacji. Zmienne ciągłe należy poddać normalizacji min-max jeśli stosowana będzie sigmoidalna funkcja aktywacji.
Neurony wyjściowe zwracają wartość z przedziału [0,1] lub [-1,1] przy stosowaniu tangensa hiperbolicznego. Sieć jest jednokierunkowa, jeśli przepływ jest dozwolony tylko w jednym kierunku, nie występują w niej pętle ani cykle. Najczęściej spotykane sieci są 3-warstwowe. Składają się z warstwy wejściowej, ukrytej i wyjściowej. Warstwowa sieć neuronowa jest pełna, gdy każdy neuron z warstwy jest połączony tylko z neuronami z warstwy następnej i nie jest połączony z żadnym neuronem ze swojej warstwy.
Warstwa wejściowa - zależy od liczby i typu zmiennych w zbiorze danych. Liczba neuronow jest

równa sumie liczby zmiennych ilościowych i łącznej liczby kategorii zmiennych jakościowych. Przekazuje ona dane do warstwy ukrytej bez przetwarzania.
Warstwa ukryta – zależy od użytkownika. Większa liczba neuronów powoduje zwiększenie mocy obliczeniowej i elastyczności sieci przy poznawaniu skomplikowanych wzorców. Zbyt duża prowadzi do przeuczenia.
Warstwa wyjściowa – zależy od zadania wykonywanego przez sieć. Przy szacowaniu mamy jeden neuron wyjściowy, przy klasyfikacji tyle, ile kategorii ma zmienna celu.

PRAWDA:

1. Może mieć tylko jedna warstwę wyjściową.
2. W przypadku klasyfikacji liczba neuronów w warstwie wyjściowej jest większa niż w przypadku szacowania.
3. Każdej zmiennej ilościowej odpowiada jeden neuron warstwy wejściowej
4. W warstwie wejściowej nie występuje funkcja aktywacji pomiędzy neuronami

FAŁSZ:

1. Uczenie sieci polega na dobieraniu odpowiedniej funkcji aktywacji w neuronach.
2. W warstwie wejściowej występuje funkcja aktywacji.
3. Uczenie sieci polega na dobieraniu liczby neuronów w warstwie ukrytej.
4. Może mieć więcej niż jedną warstwę wyjściową.

RBF

Sieci radialnej funkcji bazowej mogą być opisywane podobnym schematem jak MLP. Do warstwy wejściowej wchodzą zmienne x_1 , x_2 … x_p – predyktory. Zmienne ilościowe są normalizowane, jakościowe zamieniane są zamieniane na zmienne wskaźnikowe.
Połączenia pomiędzy warstwa wejściową, a warstwą ukrytą nie mają wag.

Warstwa ukryta – Liczba neuronów w warstwie ukrytej (zawsze jeden) jest zależna od liczby skupisk występujących w danych znajdowanych metodą dwustopniowej analizy skupień.

Różnice miedzy RBF a MLP

Zamiast sigmoidalnej funkcji aktywacji używana jest gęstość rozkładu normalnego.
Połączenia między warstwą wejściowa a ukryta nie mają wag. Zamiast tego liczona jest pozycja neuronów warstwy ukrytej.
Połączenia między warstwą ukrytą a wyjściową mają wagi. Są one optymalizowane w fazie uczenia się sieci.
Neurony wyjściowe są takie same jak w MLP. Używana jest identycznościowa bądź sigmoidalna funkcja aktywacji
RBF szybciej się uczy

PRAWDA:

1. Może mieć tylko jedną warstwę wyjściową.
2. W przypadku klasyfikacji liczba neuronów w warstwie wyjściowej jest większa niż w przypadku szacowania.
3. Współczynniki występujące we wzorach funkcji aktywacji są zależne od umiejscowienia i wielkości skupisk występujących w danych.
4. W przypadku klasyfikacji liczba neuronów w warstwie wyjściowej jest równa liczbie klas zmiennej celu.
5. Tylko połączenia pomiędzy warstwa ukryta i wyjściową mają wagi

FAŁSZ:

1. Uczenie polega na wielokrotnym korygowaniu wag przypisanych do połączeń między neuronami.
2. W warstwie wyjściowej występuje radialna funkcja aktywacji.
3. Może mieć więcej niż jedną warstwę ukrytą

Reguły asocjacyjne

Są to reguły postaci „ Jeżeli …. -> to ….” Z określonymi miarami wsparcia i ufności. Załóżmy, że mamy dwa zbiory transakcji A (np. transakcje zawierające piwo) i B (np. transakcje zawierające chipsy). Wtedy reguła asocjacyjna przybiera postać A -> B, gdzie A i B są właściwymi podzbiorami zbioru wszystkich artykułów i są rozłączne. Oznacza ona, że jeśli wśród zakupów klienta znalazło się piwo, to znalazły się wśród nich także chipsy.

Klasyfikacja reguł asocjacyjnych (ich cechy)

1. Ze względu na typ przetrzymywanych danych np.:
    a. binarne (binary lub boolean)
    b. Ilościowe (quantitative)
2. Ze względu na wymiarowość przetwarzanych danych:
    a. Jednowymiarowe
    b. Wielowymiarowe
3. Ze względu na stopien abstracji przetwarzanych danych
    a. Jednopoziomowe – jeżeli występujące w niej dane reprezentują tę samą dziedzinę wartości produkt = chipsy -> produkt = piwo
    b. Wielopoziomowe lub uogólnione jeżeli wystepujace w niej dane reprezentują różne dziedziny wartości: wiek = ’30 … 40’ -> wyksztalcenie = ‘wyzsze’

Obliczanie reguł:

Na 1000 klientów pewnego sklepu 600 kupiło napój gazowany, 180 kupiło sok, z czego 30 kupiło jednocześnie napój gazowany i sok. Dla reguły *Jeśli sok, to napój gazowany* oblicz i podaj wynik:

1. **Pokrycie:**

$$cov(X \to Y) = P(X)$$
$$P(X) = \frac{180}{1000}$$

2. **Wsparcie:** (wsparcie = pokrycie * ufność)

$$s(X \to Y) = P(X \cap Y)$$
$$P(X \cap Y) = \frac{30}{1000}$$

3. **Ufność:**

$$c(X \to Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)}$$

$$P(Y|X) = \frac{\frac{30}{1000}}{\frac{180}{1000}} = \frac{3}{100} * \frac{100}{18} = 0.1(6)$$

4. **Wzrost:**

$$I(X \to Y) = \frac{c(X \to Y)}{P(Y)}$$
$$\frac{0.1(6)}{\frac{600}{1000}} = \frac{166.(6)}{600} = 0.2(7)$$

5. **Wdrażalność –** różnica miedzy pokryciem a wsparciem.

$$P(X) - P(X \cap Y)$$
$$\frac{180}{1000} - \frac{30}{1000} = \frac{150}{1000}$$

Metody liczenia odległości miedzy grupami

Metoda pojedynczego połączenia (najbliższego sąsiada)

Odległość pomiędzy dwoma skupieniami, to odległość pomiędzy ich dwoma najbliższymi punktami. Ma tendencję do tworzenia długich i cienkich grup. Punkty osobliwe mogą prowadzić do łańcuchowania klastrów. Może dawać grupy o dużych średnicach.

Metoda całkowitego połączenia (najdalszego sąsiada)

Odległość pomiędzy dwoma skupieniami to odległość miedzy dwoma najbardziej oddalonymi ich punktami. Ma tendencję do tworzenia zwięzłych, kulistych grup. Wszystkie rekordy znajdują się wewnątrz kuli o danej średnicy. Obserwacje z danego klastra mogą być bardziej podobne do obserwacji z innego klastra niż ze swojego.

Metoda średnich grupowych (średniej odległości między grupami)

Odległość pomiędzy dwoma skupieniami to średnia odległość wszystkich rekordów z pierwszej grupy od wszystkich rekordów z drugiej grupy. Prawie równa zmienność wewnątrz grup. Prowadzi do grup bardziej podobnych kształtem do grup uzyskanych metodą całkowitego niż pojedynczego połączenia. Nie występuje zjawisko łańcuchowania klastrów. Zależna od skali w jakiej mierzone są odległości (w metodach całkowitego i pojedynczego połączenia mamy zależność tylko od porządku odległości)

Metoda centroidalnego połączenia

Odległość pomiędzy klastrami to odległość pomiędzy ich środkami. Odporne na występowanie punktów osobliwych. Nie występuje łańcuchowanie klastrów.

Metoda średniego połączenia

Odległość pomiędzy klastrami to średnia odległość pomiędzy wszystkimi rekordami łączonych grup. (także rekordami z tej samej grupy)

Metoda Warda

Liczony jest wzrost sumy kwadratów odległości rekordów od centrum przed i po połączeniu dwóch klastrów. Idea polega na zminimalizowaniu tego wzrostu w każdym kroku algorytmu.