**1. Metody eksploracji danych**

1. Co to jest eksploracja danych?

To proces poszukiwania wzorców (reguł, zależności) w dużych zbiorach danych, w celu wydobycia użytecznej wiedzy.

2. Czym są dane, informacje i wiedza? Jakie występują pomiędzy nimi zależności?

Dane – surowe, odbierane z otoczenia sygnały wyrażone w prostej postaci.

Informacja – stanowi efekt obróbki danych w taki sposób, że staje się ona użyteczna dla realizacji celów.

Wiedza – jest rezultatem przetwarzania zbiorów danych lub informacji i stanowi opis określonego zjawiska.

3. Jakie są typowe problemy rozważane na gruncie eksploracji danych?

Odkrywanie złożonych wzorców w złożonych typach danych. Praca nad eksploracją przebiegów czasowych i danych sekwencyjnych. Eksploracja danych opisujących procesy (np. przepływy pracy). Bezpieczeństwo, poufność i spójność danych.

4. Jakie są typowe metody stosowane w eksploracji danych?

Klasyfikacja wzorcowa, klasyfikacja bezwzorcowa, regresja wieloraka, sieci neuronowe, drzewa decyzyjne

5. Czym różni się analiza danych od eksploracji danych?

Analiza danych to proces przetwarzania danych w celu uzyskania na ich podstawie użytecznych informacji i wniosków. Eksploracja danych polega na znajdowaniu ukrytych prawidłowości w danych zgromadzonych w hurtowniach danych.

6. Jakie są etapy eksploracji danych w metodologii CRISP-DM?

1.Zrozumienie zasad biznesowych 2.Zrozumienie danych 3.Przygotowanie danych źródłowych 4.Modelowanie 5.Ewaluacja – ocena modelu eksploracji danych 6.Wdrożenie modelu

**2. Zastosowanie sieci neuronowych MLP w problemach regresyjnych**

1. Czym są sztuczne sieci neuronowe?

To systemy przetwarzania danych inspirowane budową i sposobem działania systemów nerwowych organizmów żywych.

2. Jak zbudowana jest sztuczna sieć neuronowa?

Składa się z warstwy wejściowej, ukrytej i wyjściowej w których znajdują się neurony.

3. Jakie są metody uczenia sieci neuronowych i czym się różnią?

Uczenie z nauczycielem (nadzorowane) -> z prezentacją kolejnego zestawu danych dostarczana jest również prawidłowa odpowiedź jaka powinna być w takim przypadku wygenerowana przez sieć.

Uczenie bez nauczyciela (nienadzorowane) -> prezentowany jest ciąg uczący, lecz nie dostarcza się sieci oczekiwanej odpowiedzi.

4. Od czego zależy liczba neuronów w warstwie wejściowej, a od czego w warstwie wyjściowej?

Liczba neuronów w warstwie wejściowej jest tyle ile jest zmiennych wejściowych. Liczba neuronów w warstwie wyjściowej zależy od typu problemu.

5. Od jakiej liczby najlepiej zacząć dobór optymalnej liczby neuronów w warstwie ukrytej?

Należy zaczynać dobór optymalnej liczby neuronów w warstwie ukrytej od małej liczby.

6. Jak działa / do czego służy funkcja aktywacji neuronu?

Funkcje aktywacji neuronu służą do transformacji poziomu aktywacji neuronu w sygnał wyjściowy.

7. Jak wyznaczane są początkowe wartości wag?

Losowo (?)

8. Czym są epoki w uczeniu sieci neuronowych?

Epoka w uczeniu sieci neuronowych to jeden ciąg uczący.

9. Jak często modyfikowane są wagi neuronów?

Wagi zmieniane są co każdy jeden ciąg uczący.

10. Czym jest wsteczna propagacja błędów?

To algorytm uczenia nadzorowanego sieci neuronowych. Zmienia wagi dowolnych połączeń elementów przetwarzających rozmieszczonych w sąsiednich warstwach sieci.

11. Jakie jest przeznaczenie zbiorów uczącego, testowego i walidacyjnego?

Zbiór uczący jest używany w celu przeprowadzania analiz na nim, a zbiory testowy i walidacyjny służą do potwierdzenia wiarygodności jej wyników. Zbiory te są efektem podziału próby statystycznej na podzbiory w sprawdzianie krzyżowym.

12. Ile razy i kiedy w uczeniu i ocenie jakości sieci neuronowej wykorzystywane są zbiory uczący,

testowy i walidacyjny?

Zbiór uczący używany jest przy modyfikacji wag, zbiór testowy przy monitorowaniu procesu uczenia, a zbiór walidacyjny przy ocenie sieci po zakończeniu uczenia.

13. Czym jest przeuczenie sieci i po czym je poznać?

Przeuczenie sieci występuje wtedy gdy model ma zbyt dużo parametrów w stosunku do rozmiaru próby na podstawie której był konstruowany. Przeuczenie można poznać, gdy model będzie dawał świetne wyniki do danych z próby uczącej, a słabe wyniki dla nowych danych, których nie było w procesie uczenia.

14. Czym jest zdolność sieci do generalizacji?

Sieć neuronowa daje dobre wyniki nie tylko w przypadku danych z próby uczącej, ale również z innych, nowych danych.

15. Kiedy należy przerwać uczenie sieci neuronowej?

Należy zastosować środki zapobiegawcze w postaci zbiorów testowych, walidacji krzyżowej lub bootstrapu, które pozwalają stwierdzić, w którym momencie dalsze uczenie może prowadzić do powstawania błędów.

16. Gdzie w sieci neuronowej zakodowane są informacje wydobyte podczas uczenia sieci?

???

17. Czym różnią się szacowanie i predykcja jako rodzaje problemów regresyjnych?

Szacowanie to przybliżone określanie wartości jakiejś wielkości przy posiadaniu niepełnych danych. Predykcja to wnioskowanie o przyszłej wartości jakiejś wielkości w określonym momencie, gdy nie jest znana wielkość wyjściowa.

**3. Zastosowanie regresji wielorakiej i sieci neuronowych RBF w problemach regresyjnych**

1. Czym jest regresja wieloraka?

Metoda statystyczna pozwalająca na badanie związku pomiędzy wielkościami danych. W regresji wielorakiej występuje więcej niż jedna zmienna objaśniająca.

2. Jak należy interpretować wartość R2 w modelu regresji wielorakiej?

Wartość R2 informuje jaka ilość regresji jest objaśniana przez stworzony model.

3. Kiedy mówimy o dużym dopasowaniu modelu (jego wysokiej jakości) w przypadku regresji

wielorakiej?

Kiedy wartość R2 jest wysoka.

4. Jak należy interpretować wartości b i b\* (β) w modelu regresji wielorakiej?

Zmienna **b** mówi o tym jaki wpływ na wartość zmiennej zależnej ma konkretna zmienna niezależna

Zmienna **b\*** określa wpływ zmiennej na wynik końcowy (dla zmiennych zestandaryzowanych)

β – wektor współczynników regresji

5. Czym różnią się współczynniki regresji wielorakiej b i b\* (β)?

**B** ma różne skale, nie można porównywać **B\*** ma jedną skalę, można porównywać

6. Co to znaczy, że współczynnik regresji jest istotny statystycznie w modelu regresji

wielorakiej?

Poziom istotności jest wynikiem, dzięki któremu można ocenić na ile badane zjawiska są ze sobą powiązane przez przypadek.

7. Czym są obserwacje odstające w modelu regresji wielorakiej?

Są to wartości które znacznie odbiegają od średniej obserwacji.

8. Jak zbudowana jest sieć neuronowa typu RBF?

Jest to siec jednokierunkowa, uczona w trybie nadzorowanym. Posiada dokładnie jedna warstwę ukrytą.

9. Jak zbudowane są neurony w warstwie ukrytej sieci RBF?

Każdy neutron warstwy ukrytej musi być połączony ze wszystkimi neuronami warstwy wejściowej. (każdy z każdym).

10. Jakie to są funkcje radialne?

Funkcje radialne to takie funkcje, które są używa jako funkcje aktywacji w sieciach neuronowych.

11. Czy sieci neuronowe RBF sprawdzają się w rozwiązywaniu problemów regresyjnych?

Tak (?)

**4. Przetwarzanie informacji jakościowych w modelach opartych na regresji wielorakiej**

1. Na jakich skalach mogą być wyrażone wyniki obserwacji i czym się te skale charakteryzują?

Skala nominalna – wartości są zbiorem symboli, które pełnią funkcję etykiet (słaba skala pomiarowa)

Skala porządkowa – wartości wskazują na przynależność do klas (słaba skala pomiarowa)

Skala Przedziałowa – własności skali porządkowej + umowne zero i stała jednostka miary

Skala ilorazowa – własności skali przedziałowej + bezwzględne zero

2. Jakie są metody reprezentowania wartości nominalnych w modelach eksploracji danych?

Metoda jeden-z-N, kodowanie równoboczne, reprezentacja oparta na pojedynczym neuronie.

3. Jakie są metody reprezentowania wartości porządkowych w modelach eksploracji danych?

Reprezentacja termometryczna I i II, reprezentacja oparta na pojedynczym neuronie.

4. Jak występowanie zmiennych jakościowych wpływa na liczbę neuronów na wejściu/wyjściu?

???

**5. Przetwarzanie informacji jakościowych w modelach opartych na sieciach neuronowych**

1. Na jakich skalach są z reguły wyrażone zmienne wyjściowe w problemach regresyjnych?

Na skalach przedziałowych i ilorazowych.

**7. i 8. Klasyfikacja wzorcowa - wprowadzenie**

1. Na czym polega klasyfikacja wzorcowa?

Klasyfikacja wzorcowa to odwzorowanie danych w zbiór predefiniowanych klas. Na podstawie danych budowany jest model, który służy do klasyfikowania nowych obiektów lub zrozumienia istniejącego podziału obiektów klasy.

2. Z jakich etapów składa się klasyfikacja wzorcowa?

Klasyfikacja wzorcowa polega na przypisywaniu obiektów do zdefiniowanych klas.

3. Jak działa metoda k-najbliższych sąsiadów?

Cały zbiór uczący jest zapamiętywany. Nowy obiekt klasyfikowany jest na podstawie znajomości klas dla jego najbliższych sąsiadów.

4. Jak działa naiwny klasyfikator Bayes’a?

Najpierw oblicza się prawdopodobieństwo a priori dla wybranej danej, a następnie szansę na to do jakiego zbioru będzie należała.

5. Jak oceniana jest jakość klasyfikacji?

Stosuje się kryteria: dokładności, efektywności, odporności modelu, skalowalności i interpretowalności

6. Na czym polega klasyfikacja z wykorzystaniem drzew klasyfikacyjnych?

Drzewo składa się z węzłów do których jest przypisany warunek. Jego spełnienie lub nie odpowiednie przenosi obiekt w dół drzewa. Liście to węzły końcowe – reprezentują odpowiednie klasy.

7. Jak wygląda algorytm budowy drzewa klasyfikacyjnego?

1. Stworzenie zbioru początkowego, i = 1 2. Stworzenie możliwych podzbiorów 3. Ocena jakości podzbiorów 4. Realizacja podziału wg kryteriów 5. Realizacja kryteriów do momentu stopu 6. Przypisanie i = i+1 i powrót do kroku 2

8. Jak zbudowane jest drzewo klasyfikacyjne?

Drzewo składa się z węzłów do których jest przypisany warunek. . Jego spełnienie lub nie odpowiednie przenosi obiekt w dół drzewa. Liście to węzły końcowe – reprezentują odpowiednie klasy.

9. W jakim celu stosowane jest przycinanie drzewa klasyfikacyjnego?

W celu upraszczania modelu i zapobiegania przeuczeniu.

10. Jakie sieci neuronowe są wykorzystywane w klasyfikacji wzorcowej?

Perceptron wielowarstwowy, RBF

**9. i 10. Klasyfikacja bezwzorcowa**

**KM**