Egzamin z Eksploracji danych dla kierunku Informatyka

Zagadnienia

1. Eksploracja danych – big data, definicja eksploracji danych, zastosowania, użyteczność, obowiązki ustawowe, zadania eksploracji danych, metodologia CRISP DM.
2. Postać i obróbka danych – zbiory danych, typy zmiennych i skale pomiarowe, postępowanie z brakami danych, normalizacja i standaryzacja, identyfikacja punktów odstających, transformacje w celu złagodzenia skośności.
3. Podstawowe statystyki: miary tendencji centralnej i rozproszenia (bez wzorów na kwantyle), wykres skrzynka z wąsami, zasada trzech sigm, skośność i kurtoza (bez wzorów).
4. Opis danych wielowymiarowych - współczynniki korelacji liniowej Pearsona i Spermana, wzór, interpretacja, siła korelacji.
5. Ocena jakości klasyfikacji i szacowania: podział na podzbiory, walidacja krzyżowa, miary jakości szacowania.
6. Ocena jakości klasyfikacji: rodzaje klasyfikacji, macierz pomyłek, wskaźniki klasyfikacji binarnej, krzywa ROC, klasyfikacja do więcej niż dwóch klas,.
7. Algorytm k najbliższych sąsiadów – zastosowania, mierzenie odległości między obiektami, przekształcanie zmiennych, zasada działania algorytmu, wybór liczby sąsiadów, szczególne sytuacje, przewidywane prawdopodobieństwa, postępowanie w przypadku zmiennej ciągłej, zalety, wady.
8. Drzewa decyzyjne – budowa i działanie, specyfika drzew CART, kryterium nieczystości węzłów i miara poprawy, zapobieganie przeuczeniu, problem zdarzeń rzadkich, postępowanie z brakami danych, ważność zmiennych.
9. Algorytm ID3 – entropia, zysk informacji, problem nadmiernej liczby klas. Algorytm C4.5 – współczynniki podziału i zysku, algorytm C5.0.
10. Bagging, boosting i lasy losowe.
11. MLP – zasada działania, schemat perceptronu, warstwy wejściowa, wyjściowa i ukryta, połączenia pomiędzy neuronami, funkcje aktywacji, uczenie sieci i problemy w uczeniu, analiza czułości.
12. RBF – zasada działania, schemat sieci, warstwy wejściowa, wyjściowa i ukryta, połączenia pomiędzy neuronami, funkcje aktywacji, uczenie sieci.
13. Regresja – definicja, model prostej regresji liniowej (wzór), metoda najmniejszych kwadratów, interpretacja współczynników modelu, współczynnik determinacji i jego interpretacja, punkty odstające, założenia modelu regresji, niebezpieczeństwo ekstrapolacji, transformacje w celu osiągnięcia liniowości.
14. Regresja wielokrotna – model, miary dopasowania, skorygowany współczynnik determinacji, współliniowość, metody wyboru zmiennych objaśniających.
15. Algorytm k średnich – cel grupowania, kroki algorytmu, wynik działania, uwagi, wady, wybór początkowych centrów skupień i ich uaktualnianie w IBM SPSS Statistics.
16. Grupowanie hierarchiczne – miary podobieństwa/niepodobieństwa dla zmiennych ciągłych, liczebności i binarnych (przynajmniej po dwie), metody łączenia grup, podstawowy hierarchiczny algorytm grupowania, dendrogramy – budowa i odczytywanie, zalety i wady.
17. Dwustopniowa analiza skupień – kroki, zalety, CF-drzewo, grupowanie klastrów, kryteria informacyjne, automatyczne wyodrębnianie klastrów, miara sylwetki.
18. Sieci Kohonena: schemat, działanie sieci, trzy charakterystyczne procesy, uczenie sieci, algorytm Kohonena, rodzaje sąsiedztwa.
19. Reguły asocjacyjne: klasyfikacja reguł asocjacyjnych, macierzowy a transakcyjny format danych, miary jakości reguł: pokrycie, wsparcie, ufność, wzrost, wdrażalność. Mocne reguły, zbiory częste, algorytm A priori.
20. Redukcja wymiaru: składowe główne, związek pomiędzy składowymi głównymi a wartościami i wektorami własnymi macierzy kowariancji, wnioski, możliwość redukcji wymiaru, kryteria określenia liczby składowych.