Systemy uczące się – Lab 2

dr Przemysław Juszczuk

Katedra Inżynierii Wiedzy, Uniwersytet Ekonomiczny

12 X 2018

- e-mail: przemyslaw.juszczuk@ue.katowice.pl
- Konsultacje: na stronie katedry + na stronie domowej
- Pokój 207 budynek A
- pjuszczuk.pl

Zaliczenie ćwiczeń – 70%

Prezentacja i demonstracja zaimplementowanego systemu wykorzystującego techniki maszynowego uczenia połączona z dyskusją na temat jego teoretycznych podstaw.

Zaliczenie wykładu – 30%

Test wielokrotnego wyboru bez możliwości korzystania z notatek.

Zadanie 1 – baza Mushroom

Na stronie UCI dostępny jest zbiór Mushroom. Zbiór opisuje zestaw próbek odpowiadających 23 rodzajom grzybów z 22 atrybutami. Wszystkie atrybuty są atrybutami kategorycznymi. W zbiorze brakuje około 2500 wartości. Należy wczytać dane, a następnie uzupełnić brakujące wartości w następujący sposób:

- wczytanie obiektu z brakującą wartością;
- wyszukanie w zbiorze danych obiektu najbardziej zbliżonego do powyższego, gdzie: wyszukujemy tylko w zbiorze, w których nie ma brakujących danych. "Najbardziej podobny" oznacza obiekt o jak największej liczbie identycznych wartości poszczególnych atrybutów;
- wartość brakującego atrybutu ustalana jest na podstawie wartości atrybutu z obiektu najbardziej podobnego.

Przykład

p,x,y,y,f,f,f,c,b,p,e,?,k,k,n,n,p,w,o,l,h,v,p
p,x,y,y,t,f,f,c,b,e,e,p,k,k,n,n,p,w,o,l,w,v,w
p,f,f,g,f,f,f,c,b,g,e,b,k,k,n,p,p,w,o,l,h,y,g
? = p

Zadanie 2 – baza Abalone

Na stronie UCI dostępny jest zbiór Abalone. Dane z pliku należy wczytać do tablicy dwuwymiarowej, a następnie przeprowadzić dyskretyzację wartości wszystkich atrybutów typu double (z wyjątkiem pierwszego atrybutu – char oraz ostatniego atrybutu – integer):

- dyskretyzacja przedziałowa atrybutów ciągłych z krokiem 0.1;
- dyskretyzacja częstościowa, gdzie liczba obiektów dzielona jest przez 10.

Zadanie 3 – baza Ecoli

Na stronie UCI dostępny jest zbiór Ecoli. Dane z pliku należy wczytać do tablicy dwuwymiarowej. Następnie przy pomocy analizy korelacji liniowej Pearsona ustalić, które atrybuty są ze sobą najsilniej skorelowane. Przeanalizować pary atrybutów: 2 i 3 oraz 6 i 7.

Zadanie 4 – baza Ecoli

Dla bazy Ecoli z zadania trzeciego dla wybranego atrybutu o wartościach typu double podzielić cały zestaw danych na 4 części: w pierwszej znajdują się dane poniżej pierwszego kwartyla, w drugiej części dane pomiędzy 1 a 2 kwartylem, w trzeciej części – dane pomiędzy 2 i 3 kwartylem. Ostatnia część to pozostałe dane – powyżej trzeciego kwartyla.