Arkadiusz Kojtek 310746

WSI

Naiwny klasyfikator Bayesa

Uruchamianie programu

Program korzysta z trzech modułów -
rand - losuje wartości które zostaną usunięte z zestawu danych. Wykorzystane one zostaną w zbiorze testującym działanie programu.
math - służy do wykonywania obliczeń matematycznych przy obliczaniu gaussowskiego rozkładu prawdopodobieństwa.
csv - służy do odczytu danych z pliku

Implementacja algorytmu

Tworzenie klasyfikatora

Do utworzenia zbioru na podstawie pliku csv służą niezależne of klasy funkcje **load_csv()** odczytująca dane z pliku oraz **divide_dataset()** dzieląca zbiór danych na zbiory przeznaczone do nauki i testów.

Program składa się z klasy **NaiveBayesClassifier** będącej implementacją naiwnego klasyfikatora Bayesa. Jako argument konstruktor przyjmuje zbiór danych. Na podstawie tego zbioru dane zostają podzielone na klasy. Następnie na podstawie słownika z klasami są obliczane:

- średnia
- odchylenie standardowe
- ilość kolumn

dla zbioru wierszy z uwzględnieniem wspomnianego podziału na klasy.

Zaklasyfikowanie nowych elementów

Gdy chcemy zaklasyfikować interesujący nas zbiór danych wywoływana zostaje metoda obiektu - **predict_rows()** przyjmująca jako argument zbiór do testowania. Metoda ta następnie dla każdego wiersza wywołuje funkcję przewidującą klasyfikacje wiersza na podstawie wcześniej przygotowanego podsumowania informacji o każdej klasie. Metoda **calculate_class_probabilites()** oblicza rozkład prawdopodobieństw na wybór każdej z klasy. Na podstawie tej informacji wybierana jest najbardziej prawdopodobna klasyfikacja.

Testy

Jako zbiór z informacjami, na którym przeprowadzane będą testy ustalony został podany na zajęciach zbiór Iris Data Set. Znajdować się on będzie w pliku iris.csv.

Do testów zbiór został podzielony w różnych, wybranych przeze mnie proporcjach.
Testy zostały przeprowadzone w proporcjach uczący do testującego odpowiednio:
1:9, 2:8, 3:7, 5:5, 7:3

Po ustaleniu proporcji losowe wiersze zestawu danych były przenoszone do odpowiednich zbiorów. Po wylosowaniu przeprowadzone zostały eksperymenty sprawdzające skuteczność klasyfikacji poprzez zliczenie ilości błędnych przypisań.
Proces losowania oraz badania został przeprowadzony dla każdej proporcji 1000 razy. Po czym została określona średnia skuteczność klasyfikacji.

Wyniki

Wyniki dla każdego badania prezentują się następująco.

Program was successful in 95.28666666666666% of cases when testing on 0.1 part of whole set
Program was successful in 95.29333333333334% of cases when testing on 0.2 part of whole set
Program was successful in 95.33444444444444% of cases when testing on 0.3 part of whole set
Program was successful in 95.30424242424243% of cases when testing on 0.5 part of whole set
Program was successful in 95.09296296296297% of cases when testing on 0.7 part of whole set

Wnioski

Jak można zauważyć naiwny klasyfikator Bayesa okazuje się wyjątkowo skuteczny. Dla badanego przez nas zbioru cechuje się skutecznością w okolicach 95%. Co jest jeszcze ciekawsze skuteczność badania pozostaje tak samo dobra niezależnie od proporcji zestawu uczącego do trenującego. Jak można zauważyć nawet w przypadku gdzie zbiór uczący stanowił tylko 0.3 całego zbioru (co daje 45 ze 150) jest to wystarczająca ilość by uzyskać bardzo dokładną klasyfikację elementów ze zbioru testującego działanie.

Naiwny klasyfikator Bayesa mimo prostej implementacji oraz surowych, i często nierzeczywistych założeń co do braku związku między poszczególnymi atrybutami różnych testowanych klas cechuje się zaskakująco dużą skutecznością ii na pewno jest godnym polecenia rozwiązaniem problemu klasyfikacji.