Paweł Galewicz 234053
Karol Podlewski 234106

Rok akademicki 2019/20
Poniedziałek, 12:30

JĘZYKI PROGRAMOWANIA W ANALIZIE DANYCH – LABORATORIUM

Zadanie 3

Opis implementacji

Stworzone rozwiązanie to program konsolowy, zaimplementowany w języku Python 3.7.5. W projekcie wykorzystano następujące biblioteki: Matplotlib, NumPy, Pandas oraz Scikit-learn.

Zbiór danych

W zadaniu wykorzystano zbiór danych zawierający informacje o zdiagnozowanych przypadkach raka piersi w stanie Wisconsin – klasyfikacja polega na przypisaniu rekordów do nowotworów złośliwych lub łagodnych. Zbiór danych zawiera 30 kolumn, wśród których znajduje się 10 cech opisanych przez trzy miary: średnią, odchylenie standardowe oraz wartości najgorsze zebrane dla każdego jądra komórkowego.

Zbiór został załadowany z paczki zbiorów biblioteki Scikit-learn, jest on możliwy do odnalezienia także w najpopularniejszych repozytoriach: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

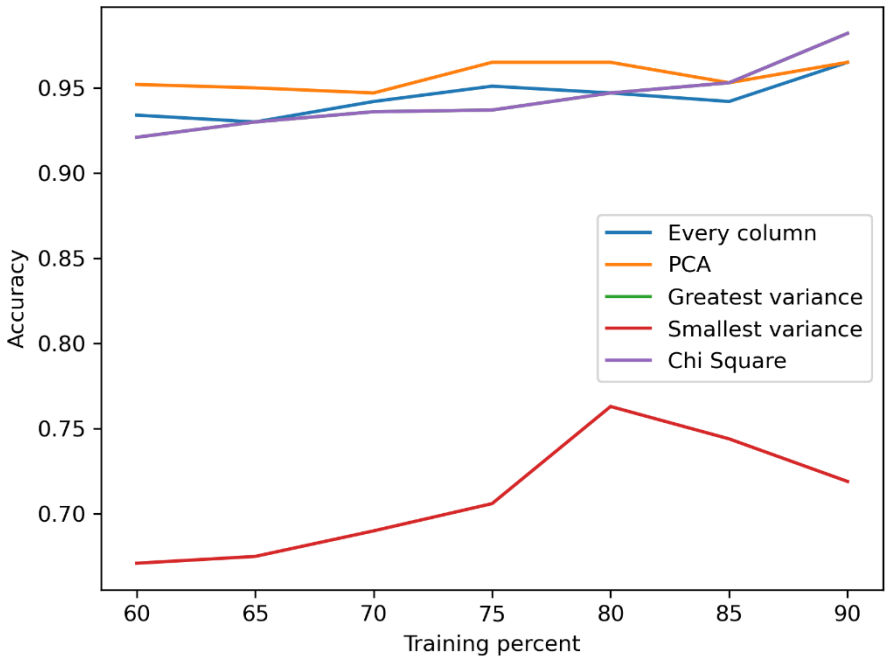
Klasyfikacja

Pierwszą częścią zadania było przeprowadzenie klasyfikacji na zbiorze danych za pomocą wybranej techniki klasyfikacji – wykorzystany został klasyfikator *maszyny wektorów nośnych*. Klasyfikację przeprowadzono na pełnym zbiorze, a także na zbiorach zredukowanych do dwóch cech z wykorzystaniem analizy głównych składowych, wyboru największej i najmniejszej wariancji oraz selekcji testem niezależności chi-kwadrat. Za miarę porównania skuteczności klasyfikacji przyjęto dokładność. Analiza została przeprowadzona dla zbiorów danych, w których część treningowa stanowiła 60%, 65%, 70%, 75%, 80% 85% lub 90% całego zbioru.

W Tabeli 1 przedstawiono cechy, do których zredukowano zbiór przy wykorzystaniu konkretnej metody. Redukcja cech przy metodzie opartej na wyborze największej wariancji oraz teście niezależności chi-kwadrat wykazały identyczne cechy, dlatego uzyskana dokładność dla każdej z metod jest identyczna – w obu wypadkach klasyfikator pracuje na identycznym zbiorze.

Tabela 1. Cechy wybrane przy redukcji cech

Metoda	Pierwsza cecha	Wartość pierwszej cechy	Druga cecha	Wartość drugiej cechy
Największa wariancja	worst area	324167,39	mean area	123843,55
Najmniejsza wariancja	fractal dimension error	0,0000070016	smoothness error	0.0000090151
Test niezależności Chi-kwadrat	worst area	112598,43	mean area	53991,66



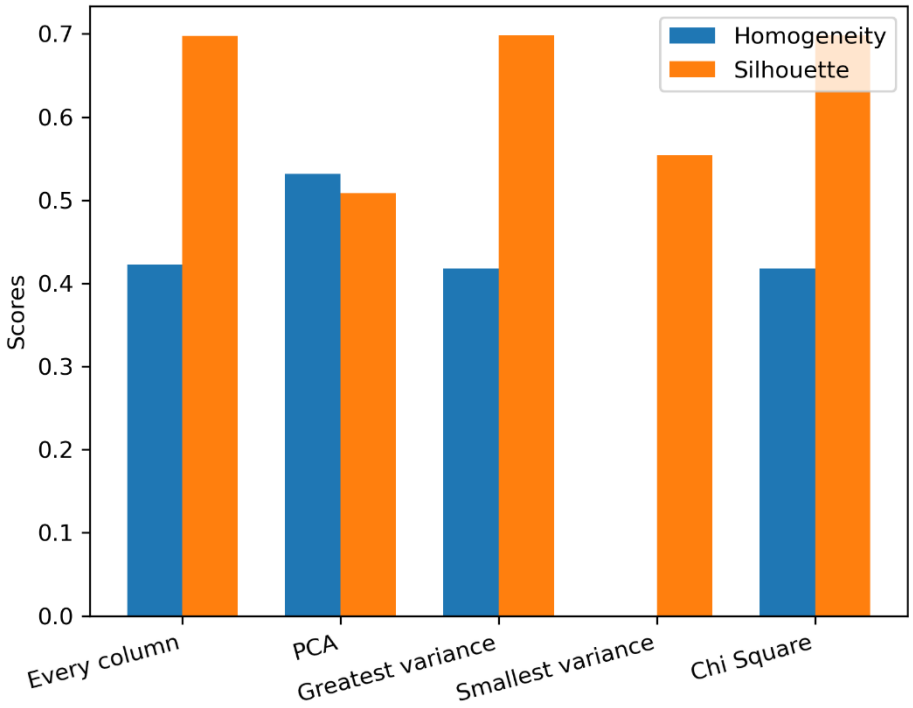
Rysunek 1. Dokładność dla klasyfikatora SVM dla pełnego oraz zredukowanych zbiorów danych

Uzyskane rezultaty pokazują, że dla zbioru opisującego zdiagnozowane przypadki raka piersi redukcja cech z wykorzystaniem najmniejszej wariancji znacząco pogarsza rezultaty klasyfikacji. Redukcja metodą analizy głównych składowych zawsze osiąga przynajmniej identyczną dokładność jak klasyfikacja na pełnym zbiorze – zazwyczaj wyniki są lepsze. Wybór cech z wykorzystaniem testu niezależności chi-kwadrat oraz wyboru cech o największej wariancji uzyskały bardzo podobną dokładność jak analiza pełnego zbioru. Przy tych metodach redukcji dokładność rosła wraz ze wzrostem zbioru treningowego, co nie charakteryzowało pracy zbiorach zredukowanych za pomocą innych metod, a także pełnym zbiorze.

Analiza skupień

Druga część zadania zakładała przeprowadzenie analizy skupień na zbiorach przygotowanych w części pierwszej. Badania przeprowadzono przy użyciu algorytmu *k-średnich*. Znając specyfikę zbiorów – dane dzielą na dwie klasy – metodę skonfigurowano tak, by wyszukiwała dwóch klastrów. Jako kryteria porównawcze przyjęte zostały dwie miary: *silhouette* oraz *jednorodność* (ang. *Homogeneity*).

Rysunek 2 prezentuje porównanie wyników miar dla rozpatrywanych zbiorów. Dla obu miar wyższy wynik interpretować można jako trafniejsze wymodelowanie skupisk do rzeczywistych klas obserwacji.



Rysunek 2. Wartości miar homogenity oraz silhouette dla k-średnich

Jednorodność wyznaczonych skupisk w każdym przypadku była na średnim poziomie – wyznaczone skupiska zawierają obserwację nowotworów i złośliwych, i łagodnych. Algorytmowi nie udało się wyznaczyć skupisk, które jednoznacznie określałyby klasę obserwacji. Świadczyć to może o tym, że dane nie są pogrupowane w proste, geometryczne kształty, z którymi *k-średnich* radzi sobie najlepiej. Zbiór z najmniejszą wariancją osiągnął wynik bliski zeru – może to być związane z faktem, że dane są rozmieszczone blisko siebie przez mały zbiór wartości. Na podstawie wartości miary silhouette stwierdzić można, że skupiska nie nachodziły na siebie, ponieważ obserwacje z jednego klastra dobrze pasowały do siebie i źle do obserwacji oraz drugiego klastra.