Zbigniew Nowacki 234102
Bartosz Jurczewski 234067

Rok akademicki 2019/20
Poniedziałek, 12:30

JĘZYKI PROGRAMOWANIA W ANALIZIE DANYCH – LABORATORIUM

Zadanie 3

Opis implementacji

Zadanie zostało zrealizowane przy użyciu języka Python w wersji 3.8.2, z wykorzystaniem bibliotek: *pandas*, *matplotlib*, *numpy* oraz *sklearn*.

Zbiór danych

Wybranym przez nas zbiorem zadań był zbiór „Leaf Classification".

Zestaw danych zawiera 990 rekordów opisujących okazy liści na podstawie wykonanych zdjęć.

Zbiór zawiera trzy zestawy cech na obraz:

- ciągły deskryptor kształtu (ang. *a shape contiguous descriptor*, kolumny **shape_X**),
- histogram tekstury wewnętrznej (ang. *interior texture histogram*, kolumny **texture_X**),
- histogram marginesu o małej skali (ang. *fine-scale margin histogram*, kolumny **margin_X**),

gdzie x to liczba od 1 do 64.

Dla każdej cechy podano 64-atrybutowy wektor na próbkę liścia.

Zbiór dostępny jest pod następującym adresem: https://www.kaggle.com/c/leaf-classification/data

Klasyfikacja

Do zrealizowania pierwszej części zadania wykorzystaliśmy naiwny klasyfikator Bayesa. Podczas eksperymentów zestawiliśmy ze sobą skuteczność (miara dokładności) oraz procent treningowy (cześć zbioru użyta jako zbiór treningowy).

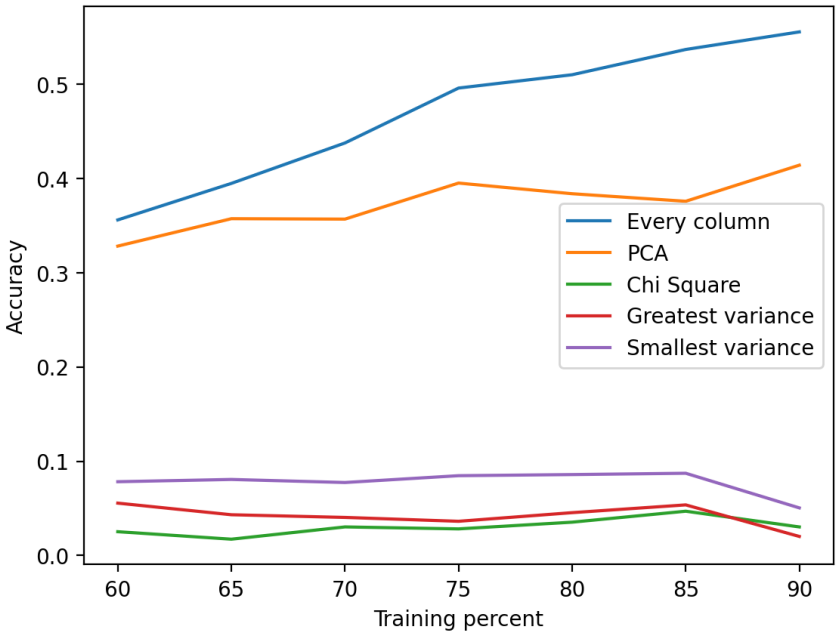
Klasyfikacja została przeprowadzona dla następujących wartości części treningowej: 60%, 65%, 70%, 75%, 80%, 85% oraz 90% całego zbioru.

Do klasyfikacji został użyty pełen zbiór, a także zbiory zredukowane do dwóch cech z wykorzystaniem analizy głównych składowych, wyboru największej i najmniejszej wariancji oraz selekcji testem niezależności chi-kwadrat.

Tabela 1. Cechy wybrane przy redukcji cech

Metoda	Pierwsza cecha	Wartość pierwszej cechy	Druga cecha	Wartość drugiej cechy
Największa wariancja	texture12	0.005	texture15	0.004
Najmniejsza wariancja	shape38	0,00000005737	shape37	0,00000005842
Test niezależności Chi-kwadrat	texture15	203.187	texture60	166.841

Tabela 1 przedstawia cechy, do których został zredukowany zbiór danych przy wykorzystaniu danej metody.



Rysunek 1. Dokładność dla naiwnego klasyfikatora Bayesa, dla zbioru danych pełnego i zredukowanych

Otrzymane przez nas wyniki pokazują, że dla naszego zbioru redukcja cech z wykorzystaniem którejkolwiek zaimplementowanej metody wpływa **negatywnie** na rezultaty klasyfikacji.

Redukcja metodą analizy głównych składowych wpłynęła najmniej negatywnie spośród wybranych metod redukcji cech. Jej wynik, w porównaniu do braku redukcji cech, pogarszał się wprost proporcjonalnie do zwiększającego się procentu danych treningowych.

Metody najmniejszej jak i największej wariacji oraz testu niezależności chi-kwadrat poradziły sobie najgorzej uzyskując dokładność mniejszą niż 0.1. Przy przeznaczeniu 90% zbioru na trening było widać zjawisko przeuczenia.

Największą dokładność naiwny klasyfikator Bayesa osiągnął przy pracy na zbiorze niezredukowanym. Może to być spowodowane tym, że zwykle działa on najlepiej przy większej liczbie cech oraz zakłada, że cechy są od siebie niezależne.