|  |  |
| --- | --- |
| 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 około 1584 zdjęć okazów liści (16 próbek po 99 gatunków), które zostały przekonwertowane na binarne czarne liście na białym tle.

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 (czyli jaka cześć zbioru posłużyła za zbiór treningowy).

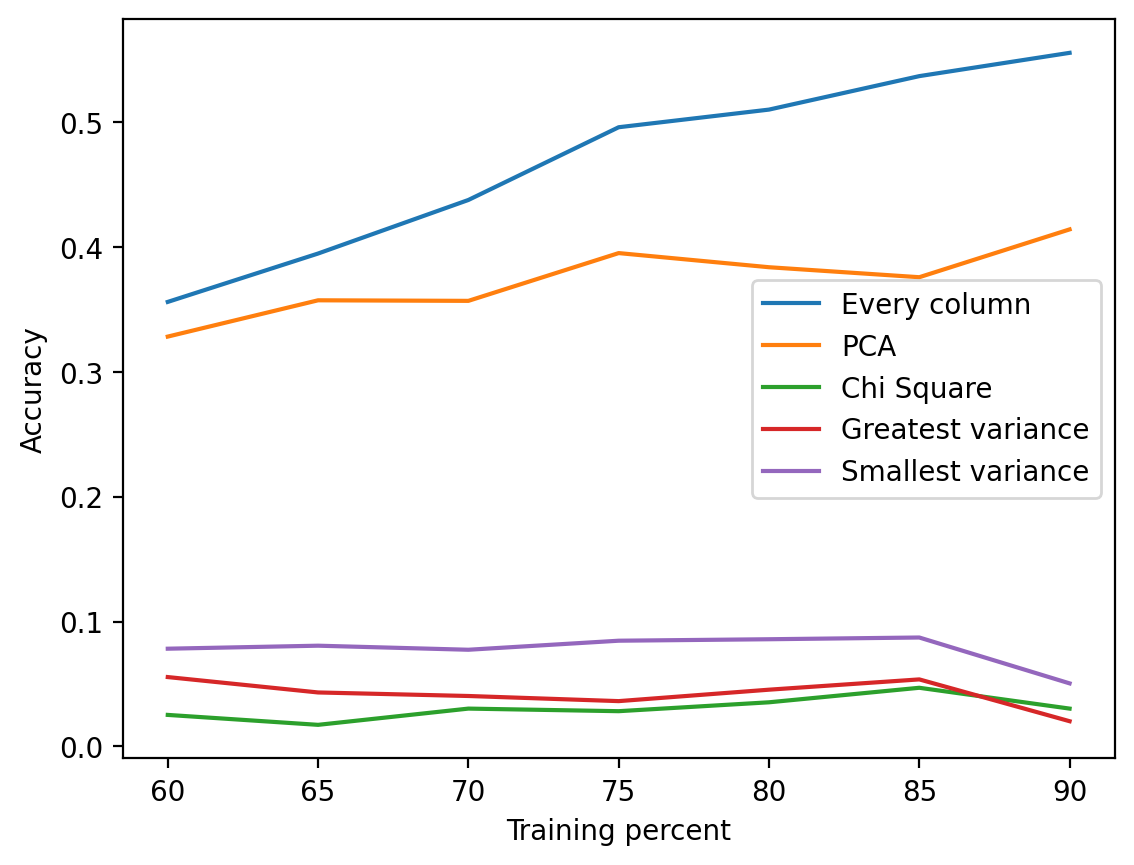
Klasyfikacja została przeprowadzona dla następujących wartości części treningowej: 60%, 65%, 70%, 75%, 80%, 85% oraz 90% całego zbioru.

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 | 5.737e-8 | shape37 | 5.842e-8 |
| 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 pełnego oraz zredukowanego zbiorów danych

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, jednakże jej wynik 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.

Największą dokładność klasyfikator osiągnął przy pracy na zbiorze niezredukowanym, co pozwala nam dojść do wniosku, że przy pracy na tym zbiorze danych, naiwnemu klasyfikatorowi Bayesa nie należy przeszkadzać.