WSI

Ćwiczenie 4 – regresja i klasyfikacja

Prowadzący: mgr inż. Mikołaj Markiewicz

Wykonał: Jan Kaniuka

Numer indeksu: 303762

Treść zadania – zaimplementować naiwny klasyfikator Bayesa (gaussowski)

Założenia – Do eksperymentów wykorzystano zbiór danych dot. jakości wina (https://archive.ics.uci.edu/ml/datasets/Wine+Quality). Eksperymenty prowadzono na zbiorze ***winequality-red.csv***.

Raport z przeprowadzonych eksperymentów:

Do weryfikacji jakości modelu wykorzystano:

- *k-krotną* walidację krzyżową (*k=5*)

- podział na zbiór treningowy i testowy (stosunek liczności zbiorów: *60/40*)

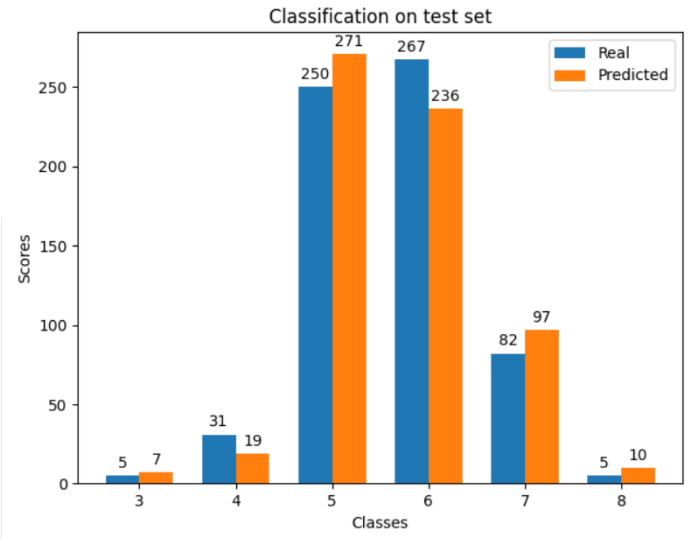
Miary oceny jakości modelu klasyfikatora

Jako miarę oceny jakości modelu przyjęto wskaźnik $accuracy = \frac{liczba\ \boldsymbol{poprawnych}\ dopasowań}{liczba\ \boldsymbol{wszystkich}\ dopasowań} \times 100\%$

5-krotna walidacja krzyżowa

Numer podzbioru testującego	1	2	3	4	5	**Średnia wartość *accuracy (%)***
accuracy (%)	56.88	59.69	55.94	53.13	52.66	**55.66**

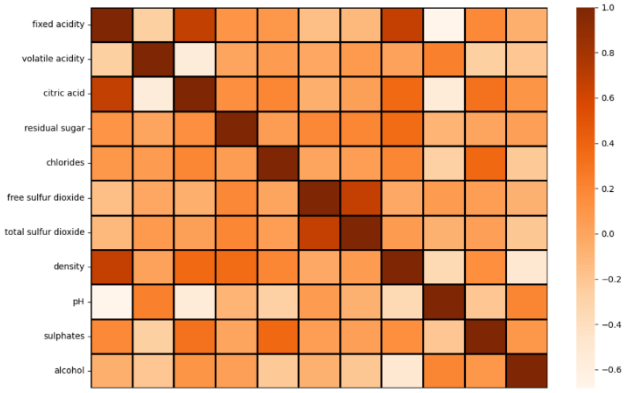
Podział na zbiór treningowy i testowy



Wartość błędu uzyskana dla klasyfikatora nauczonego na 60% zbioru i testowanego na pozostałej części zbioru wyniosła **52.1875%**.

Wykres po lewej stronie pokazuje rzeczywistą liczbę win w danych klasach w porównaniu do liczby win zaklasyfikowanych do danej klasy. Niereprezentatywna liczba win klasy 8 spowodowała, że zaklasyfikowano tam 2x więcej win niż jest ich w rzeczywistości.

Niezbyt wysoka dokładność klasyfikatora $(accuracy)$ skłoniła mnie do poszukania przyczyny, która za tym stoi. W naiwnym klasyfikatorze Bayesa zakładamy (*naiwnie* ☺) niezależność zmiennych. Wyliczyłem więc w osobnym programie ***macierz korelacji atrybutów wejściowych***. Zaobserwowałem jedynie pojedyncze, silniejsze korelacje między kilkoma zmiennymi. Wyznaczona macierz wyraźnie pokazuje, że znaczna większość zmiennych jest słabo skorelowana, więc nie jest to najprawdopodobniej przyczyna przeciętnego działania klasyfikatora.



Odpowiedź znalałem w szczegółowym opisie danych dostępnym na stronie podanej we wstępie. **Zbiór nie jest zbalansowany** i win średnich jest wielokrotnie więcej niż win bardzo dobrych lub bardzo słabych. Z tego właśnie powodu klasyfikator nie był w stanie osiągnąć wysokiej wartości wskaźnika dopasowania.

Odpowiedzi na pytania:

1) *Jakiego podzbioru danych (z tych którymi dysponujemy) użyjemy do zbudowania docelowego modelu na potrzeby klasyfikowania nowych próbek (czyli dla tych dla których budujemy klasyfikator)?*

   Odpowiedź: Użyjemy całego dostępnego zbioru danych, ponieważ wpłynie to pozytywnie na generalizację. *Overfitting* jest mało prawdopodobny, gdyż zbiór jest silnie niezbalansowany.

2) *Jak zinterpretować różnice/brak różnic w wynikach z weryfikacji jakości modelu obu metod (k-krotna walidacja vs zbiór treningowy i testowy)*

   Odpowiedź: Brak różnic (nieznaczne różnice na poziomie kilku procent) wskazuje, że jakość klasyfikatora nie zależy (w tym przypadku) od wyboru podzbioru testującego.

Wnioski dot. działania klasyfikatora:

1) Klasyfikator działa przeciętnie z powodu niezróżnicowanego zbioru dostarczonych danych.

2) W celu polepszenia działania należy zebrać więcej pomiarów (głównie dla skrajnych klas) oraz dla klas których nie uwzględniono w podanym zbiorze (brakuje klas 1,2,9,10).

3) Po rozszerzeniu bazy danych można również rozważyć dobór *lepszych/bardziej miarodajnych* wskaźników jakości. Przyjęty wskaźnik jakości na równi „karze” zaklasyfikowanie wina słabego jako bardzo dobrego oraz wina średniego jako wina średniego z sąsiedniej klasy. Oczywistym jest, że druga pomyłka ma zdecydowanie mniejszą „wagę” aniżeli ta pierwsza.