Piotr Kosmala, 310 754

WSI – ćwiczenie 7

Modele Bayesowskie

# Opis zadania

W celu wykonania ćwiczenia zaimplementowałem naiwny klasyfikator Bayesa, a następnie wykorzystałem go do rozpoznawania klas w zbiorze danych wine z pakietu scikit-learn. Do obliczania prawdopodobieństw został przeze mnie wykorzystany rozkład normalny. Powstały model został przetestowany na jednokrotnie wybranym losowym zbiorze testowym, jak również za pomocą  
n-krotnej walidacji krzyżowej.

# Testowanie

Testy przebiegały poprzez stworzenie modelu na podstawie zbioru uczącego, a następnie porównaniu przewidywanych wyników z faktycznymi znajdującymi się w zbiorze testującym. Podane wartości dokładności to stosunek prawidłowych przewidywań do liczności zbioru testowego.

## Losowy zbiór testowy

Stosunek zbioru testowego do uczącego: 2:8

Osiągnięta dokładność: 0,527

Dla dużego zbioru uczącego model osiąga dość dobrą dokładność, zbliżoną do 50%. Nie wiemy jak duża w tym przypadku jest skala overfittingu.

Stosunek zbioru testowego do uczącego: 5:5

Osiągnięta dokładność: 0,415

Wraz z obniżaniem rozmiaru zbioru uczącego spada także dokładność modelu.

Stosunek zbioru testowego do uczącego: 8:2

Osiągnięta dokładność: 0,328

Przy zbyt małym zbiorze uczącym model nie osiąga zadowalającej nas dokładności. Prowadzi to do wniosku, że naiwny klasyfikator Bayesa potrzebuje dużej ilości danych do uzyskania wystarczającej precyzji.

Powyżej przedstawiłem przykładowe wyniki wykonanych testów, jednakże potrafiły się one bardzo różnić w każdym wykonaniu programu, co wynika z losowego wyboru zbioru uczącego i testowego. Podjąłem próbę zapobiegnięcia temu poprzez zastosowanie n-krotnej walidacji krzyżowej.

## N-krotna walidacja krzyżowa

Ten sposób weryfikacji polega na n-krotnym podziale danych na zbiór uczący i testowy, dzięki czemu powinniśmy zmniejszyć różnorodność uzyskanych wyników, co wynika z faktu, że wykorzystywany jest cały zbiór danych w różnych konfiguracjach.

|  |  |  |  |
| --- | --- | --- | --- |
| n-krotność | 3 | 4 | 5 |
| dokładność | 0,668 | 0,691 | 0,679 |

Eksperyment został przeprowadzony dla podziału na 3, 4 i 5 zbiorów. Wyniki dla wszystkich sytuacji  
są jednak porównywalne, zatem można potraktować je wspólnie. W toku testowania widoczne okazało się, że kolejne wykonania programu nie wpływają znacznie na uzyskiwane wartości, które cały czas znajdowały się bardzo blisko 67%.

# Wnioski

Naiwny klasyfikator Bayesa osiągnął dokładność bliską 67% na zbiorze danych wine.  
Wynika to prawdopodobnie z niewielkiego rozmiaru tego zestawu, który sprawił, że tworzony model nie mógł być wystarczająco dokładny. Teza ta wydaje się uzasadniona na mocy eksperymentów  
z różnymi licznościami zbioru uczącego, które pokazały, że implementowany algorytm wymaga dużej ilości danych dla uzyskania zadowalających efektów.

N-krotna walidacja krzyżowa okazała się bardzo skuteczną metodą weryfikacji dokładności modelu, poprzez eliminację nadmiernego wpływu losowości doboru zbioru testowego. Zachowuje się ona zdecydowanie lepiej niż używany do tej pory przeze mnie sposób mierzenia parametrów implementowanych algorytmów uczenia maszynowego.