

Regresja i klasyfikacja

Zadanie:

Zaimplementować naiwny klasyfikator Bayesa (Gaussowski). Zweryfikować jakość modelu za pomocą k -krotnej walidacji krzyżowej oraz podziału na zbiór treningowy i testowy(60/40).

Do eksperymentów został wykorzystany zbiór danych o jakości białego wina. Jako kryterium skuteczności działania klasyfikatora wybrałem jego precyzyjność, czyli ilość poprawnie ocenionych jakości win ze zbioru testowego podzielonych przez ilość wszystkich win z owego zbioru. W przypadku walidacji krzyżowej liczyłem precyzję w każdej iteracji oraz średnią wartość tych precyzji. Trafność takiej miary oceny bardzo zależy od jakości zbioru uczącego(na ile równomierny jest rozkład różnej jakości win), natomiast z drugiej strony jest on bardzo prosty i intuicyjny

Wyniki eksperymentów:

W celu ulepszenia jakości przeprowadzanych eksperymentów dane w zbiorze były losowo przetasowywane.

1. Podział na zbiór testowy i treningowy:

Przetestowałem działanie klasyfikatora kilkakrotnie w celu otrzymania różnych zbiorów treningowych i odpowiednio różnej jakości modele.

Otrzymywałem następujące wartości jakości klasyfikatora:

0.4061	0.4031	0.4122	0.4163	0.4087	0.4224	0.4199	0.4082	0.4154	0.4087
--------	--------	--------	--------	--------	--------	--------	--------	--------	--------

Średnia z tych wartości wynosi **0.4121**

2. k -krotna walidacja krzyżowa($k = 5$):

W tym przypadku jakość była oceniana analogicznie jak powyżej za wyjątkiem tego że za finalną jakość przyjęto średnią arytmetyczną ze wszystkich modeli

Iter.	1	2	3	4	5	Średnia
Jakość	0.4259	0.4147	0.4311	0.4311	0.3749	0.4155

Przeprowadziłem również eksperymenty dla różnych wartości k . W wyniku eksperymentów nie ujawniono zależności pomiędzy jakości modelu oraz ilością podziałów zbioru.

Wnioski:

Z obydwu metod dostałem prawie taką samą jakość modelu na poziomie 40%, co nie jest zaskakująco dużym wynikiem. Po sprawdzeniu ilości win o różnej jakości w naszym zbiorze:

Jakość	Ilość
6	2198
5	1457
7	880
8	175
4	163
3	20
9	5

można zrobić wniosek że ten zbiór nie jest idealnym przypadkiem na zastosowanie wybranej miary jakości, ponieważ dane są słabo zróżnicowane, większość win ma jakość 5 albo 6, natomiast win bardzo dobrych albo bardzo złych jest znacząco mniej.

Korzystając z bardziej skomplikowanych metod oceny jakości modelu dostalibyśmy bardziej dokładny wskaźnik jej skuteczności, natomiast model nadal pozostaje niedokładny.

Odpowiedzi na pytania:

- *Jakiego podzbioru danych (z tych którymi dysponujemy) użyjemy do zbudowania docelowego modelu na potrzeby klasyfikowania nowych próbek (czyli dla tych dla których budujemy klasyfikator)?*
Najlepiej by było skorzystać z całego zbioru danych którymi dysponujemy, walidacja krzyżowa pokazuje że zastosowanie modelu na każdym z 5 podzbiorów daje prawie taką samą jakość
- *Jak zinterpretować różnice/brak różnic w wynikach z weryfikacji jakości modelu obu metod (k-krotna walidacja vs zbiór treningowy i testowy)*
Wyniki walidacji różnią się nieznaczająco, natomiast wynik z walidacji krzyżowej jest bardziej stabilny i dokładny. To że obydwie metody dają prawie taki sam wynik może świadczyć o złej jakości naszego modelu