06.12.2021r

Politechnika Warszawska

Wprowadzenie do sztucznej inteligencji

Ćwiczenie 4

Zaimplementować naiwny klasyfikator Bayesa (Gaussowski) dla zadania klasyfikacji jakości białego wina.

Przemysław Krasnodębski

Link do repozytorium: [WSI-21Z/Cwiczenie 4 at master · p-krasnodebski/WSI-21Z (github.com)]

Jako miarę jakości modelu klasyfikatora wybrano współczynnik błędnych dopasowań do liczby wszystkich dopasowani(*1 – accurancy*). Wyznaczenie tej miary jakości jest stosunkowo łatwe, nie zwiększa skomplikowania obliczeniowego zadania a jednocześnie bardzo dobrze pokazuje jak dobrze model radzi sobie z klasyfikowaniem danych. W trakcie doświadczeń próbowano wyznaczyć optymalne parametry, tak aby minimalizować wspomniany współczynnik.

W poniższych doświadczeniach miarę jakości wyznaczano jako średnią z miar jakości dla 5 wywołań klasyfikatora.

Weryfikacja jakości modelu:

Zbiór treningowy i testowy	
Współczynnik podziału	Wyniki
0.1	0.5678
0.2	0.5625
0.3	0.5528
0.4	0.5439
0.5	0.5519
0.6	0.5544
0.7	0.5378
0.8	0.5363
0.9	0.5419

k-krotna walidacja krzyżowa	
k	Wyniki
2	0.5578
3	0.558
4	0.56
5	0.557
6	0.5592
7	0.557
10	0.5586
15	0.5571
20	0.5572

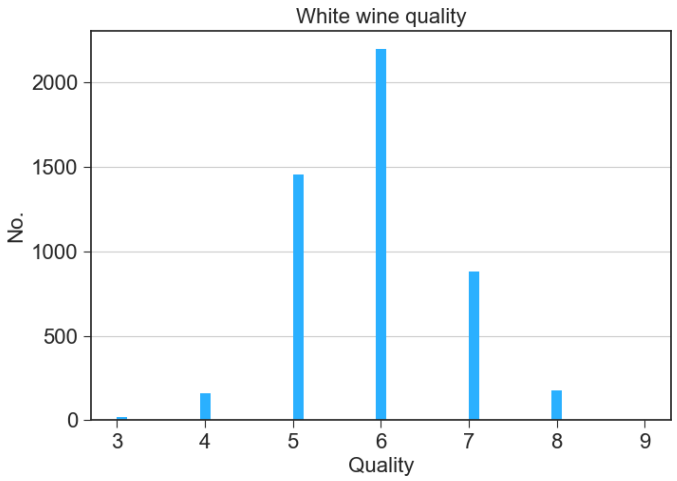
Wyniki wskazują, że algorytm działa, odpowiedni dobór parametrów poprawia jakość klasyfikatora, choć uzyskiwane błędy dopasowań są dość duże. Wyniki dla obu testowanych metod są podobne.

Eksperymentalnie spróbowano wyznaczyć wpływ poszczególnych atrybutów na proces klasyfikacji jakości białego wina. Wykorzystano 5-krotną walidację krzyżową oraz średnią miarę jakości z 5 testów. W poniższej tabeli przedstawiono miary jakości dla wywołań klasyfikatora, tylko na podstawie dwóch atrybutów.

Z tabeli jasno wynika, że nie wszystkie atrybuty jednakowo przekładają się na jakość wina. Atrybut nr 10 wydaje się najlepszy. W ramach testów łączono atrybuty o najlepszych wynikach, ale atrybut **1** i **10** dają najlepszy model, podobne wyniki zwraca też zestaw atrybutów 0, 1, 2, 10.

1	2	3	4	5	6	7	8	9	10	Atrybut
0.5412	0.5471	0.5503	0.5516	0.5559	0.5517	0.5555	0.5549	0.5542	0.5133	0
	0.5351	0.5381	0.5459	0.5405	0.5312	0.5202	0.5436	0.5425	0.4752	1
		0.5452	0.5973	0.5479	0.5404	0.5354	0.5473	0.5493	0.507	2
			0.5769	0.5541	0.5497	0.5568	0.5552	0.5536	0.5232	3
				0.5575	0.5818	0.5705	0.5557	0.5571	0.5212	4
					0.5503	0.5536	0.5575	0.5558	0.5103	5
						0.5385	0.5475	0.5507	0.5189	6
							0.545	0.5414	0.533	7
								0.5563	0.5017	8
									0.5107	9

Są to wyniki najniższe z dotychczas uzyskanych, lecz ich wartości ciągle są zbyt duże. Z tego powodu postanowiono sprawdzić jakość danych wejściowych.



Niestety dane mają niesymetryczny rozkład klas, praktycznie dla 4 klas jakości dane są zbyt ubogie. Dodatkowo klasa jakości o wartości 6 ma znacznie więcej wystąpień w zbiorze danych. Połączenie powyższych problemów znacznie wpływa na jakość klasyfikacji danych.

Pomimo złej jakości danych naiwny klasyfikator Bayesa potrafi klasyfikować dane. Skutecznie może być stosowany w wielu łatwych problemach uczenia maszynowego. Jego wadą jest założenie o wzajemnej niezależności atrybutów, lecz założenie to w wielu przypadkach nie generuje zbyt wielu problemów i jest dobrym uproszczeniem. Kolejną wadą jest problem z błędną klasyfikacją klas o których model testowy wie dość mało, można rozwiązać ten problem poprzez poszerzenie zbioru danych.

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)?

Dane wykorzystywane w modelu docelowym powinny być jak najbardziej zróżnicowane. W danym przypadku warto użyć całego zbioru danych. Wyniki uzyskane w wielu próbach(losowych i walidacji) są podobne, co oznacza, że pełny zbiór jest dobrym wyborem.

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

Brak różnic można interpretować jako fakt, iż obie te metody równie dobrze(lub równie źle) radzą sobie z weryfikacją jakości tak prostego modelu jak klasyfikator jakości wina.