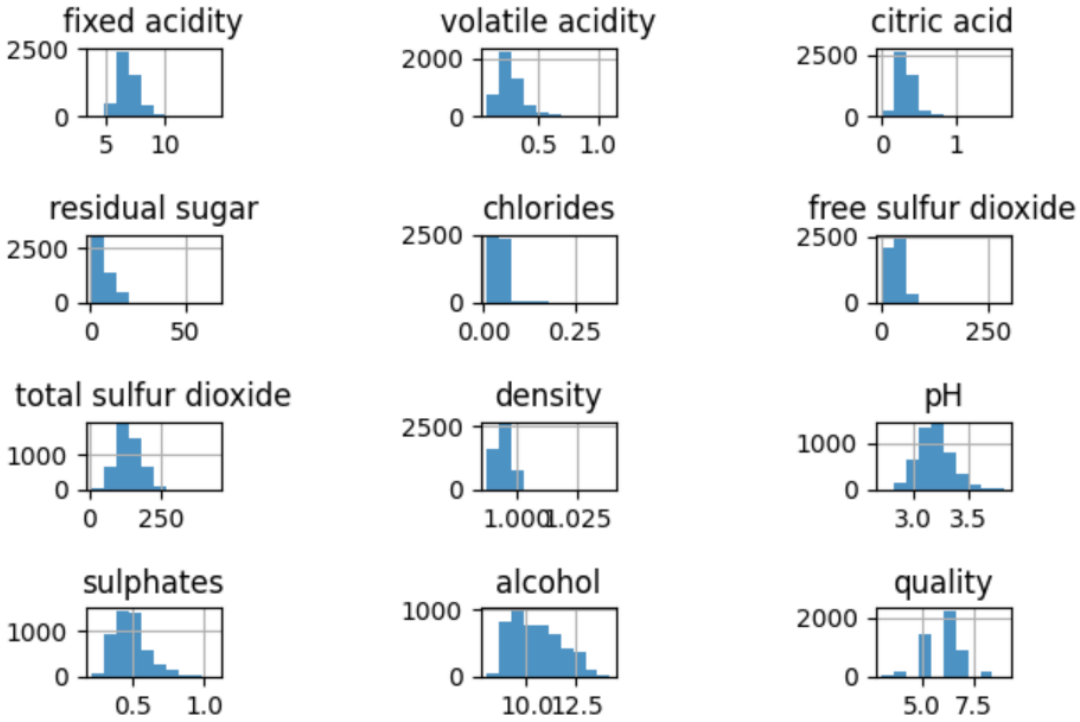
Lena Zubik, nr indeksu 303789

Temat: Naiwny klasyfikator Bayesa (Gaussowski) na przykładzie zbioru dot. jakości wina (białe)

Naiwny klasyfikator Bayesa jest przykładem klasyfikatora probabilistycznego. Dla danych numerycznych obliczamy średnią i odchylenie standardowe każdego zdefiniowanego atrybutu. Klasyfikacja bazuje na obliczonej średniej i odchyleniu standardowym obliczając prawdopodobieństwo na podstawie wartości zawracanej przez gaussowską funkcję przynależności (zakłada, że atrybuty są wzajemnie niezależne). Wybierana jest klasa, która po obliczeniu zgodnie ze wzorem Bayesa, która da największe prawdopodobieństwo. Tak nieskomplikowany klasyfikator jest często wykorzystywany ze względu na swoje niezłe rezultaty zarówno dla danych numerycznych jak i dyskretnych.

K-krotna walidacja krzyżowa polega na podzieleniu zbioru danych na k zbiorów i zwracaniu uśrednionej miary oceny jakości modeli dla przypadków, kiedy zbiór testowy jest kolejno 1-szym, 2-im, … k-tym z wydzielonych zbiorów. Jest to dobry sposób na sprawdzenie, jak rzeczywiście model radzi sobie na różnych danych.

Zobrazowanie danych



Jak widać zbiór próbek zawiera najwięcej przykładów klasy ósmej. Jest ich dużo więcej w porównaniu do ilości próbek klasy trzeciej. Zaimplementowany klasyfikator Bayesa bazuje na ocenie dopasowania poszczególnych atrybutów rozkładem Gaussa. Z załączonych histogramów jasno wynika, że niektóre dane nie dają się aproksymować takim rozkładem np. ‘chlorides’.

Wyniki przeprowadzonych eksperymentów

Jako miarę oceny jakości modelu klasyfikatora przyjęłam jego cenność oceny, czyli ilość dobrze sklasyfikowanych przypadków podzielnych przez rozmiar zbioru testowego. Uważam, że jest to intuicyjna miara oceny jakości i dokładnie odwzorowująca, to czego chcemy uzyskać od zabudowanego modelu, czyli to żeby dawał jak najczęściej poprawną klasyfikację.

Podział zbioru	30:70	40:60	50:50	60:40	70:30	80:20
Accuracy	37.8%	36.4%	39.1%	39.5%	38.8%	36.4%

W powyższej tabeli znajduje się celność modelu w zależności od wielkości zbioru trenującego i zbioru testowego. Podział zbioru jest podany w konwencji: pierwsza liczba to procent zbioru danych, jaki będzie pełnił funkcję zbioru trenującego, analogicznie druga liczba dotyczy zbioru testowego.

Stopień podziału danych	Średnia accuracy modelu	Accuracy danego zbioru
k = 2	41.4%	43.7%
		39.1%
k = 3	40.5%	44.7%
		39.3%
		37.3%
k = 4	38.7%	43.2%
		35.9%
		39.8%
		35.9%
k = 5	36.7%	36.1%
		35.2%
		37.4%
		38.5%
		36.4%

W powyższej tabeli znajdują się wyniki klasyfikacji zbudowanego modelu przy weryfikacji k-krotną walidacją krzyżową.

Z przeprowadzonych eksperymentów wynika, że najlepsze wyniki trafności modelu uzyskujemy, gdy zbiór treningowy wynosi 50%, 60% zbioru danych. Walidacja krzyżowa potwierdza, że przy rozmiarze zbioru treningowego 80% danych, model klasyfikuje gorzej, zachodzi zjawisko nadmiernego dopasowania. Z tego powodu wynik dla testu podziału zbioru 60:40 jest lepszy od walidacji krzyżowej dla k=5.

Walidacja krzyżowa natomiast dostarcza nam bardziej „obiektywnego" spojrzenia na uzyskany wynik (celność modelu) jako średnia uzyskanych rezultatów, czego dobrym przykładem jest podział zbioru danych dokładnie w połowie na zbiór trenujący i testowy. Okazuje się, że konfiguracja, w której to drugie zbiór jest trenujący daje lepszy rezultat niż ta, kiedy pierwszy jest zbiorem trenującym.

Jakiego podzbioru danych (z tych którymi dysponujemy) użyjemy do zbudowania docelowego modelu?

Do zbudowania docelowego modelu danych użyłabym podzbioru danych bliżej rozmiaru 60% zbioru danych niż 80% zbioru danych, aby uniknąć zjawiska nadmiernego dopasowania modelu do danych trenujących. Jeżeli chodzi o model z walidacją krzyżową to można użyć wszystkich próbek.

Jak zinterpretować różnice/brak różnic w wynikach z weryfikacji modelu obu metod?

Różnice między wynikami zwracanymi przez weryfikację modelu poprzez prosty podział zbiory danych na dwa zbiory a walidacją krzyżową są widoczne. Podział zbioru jest ściśle związany z danymi, które się w nim znajdą, natomiast k-krotna walidacja uśredni wyniki z każdej możliwej kombinacji takich podzbiorów, zatem uzyskany będzie bardziej obiektywny.