WSI Modele bayesowskie sprawozdanie

Małgorzata Grzanka

10.06.2024

Spis treści

1 Wstep

Celem ćwiczenia było zaimplementowanie naiwnego klasyfikatora Bayes'a. Nastepnie, napisany klasyfikator należało przetestować dla zbioru danych Breast Cancer Wisconsin Dataset (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.htmlsklearn.datasets.load_breast_cancer).

2 Zasada działania naiwnego klasyfikatora Bayes'a

2.1 Sposób klasyfikacji

Podstawa naiwnego klasyfikatora Bayes'a jest wzór na prawdopodobieństwo warunkowe Bayes'a:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Wzór ten jest wykorzystywany w naiwnym klasyfikatorze Bayes'a do przewidzenia klasy próbki pod warunkiem jej cech (B odpowiada klasie próbki, a A

- wektorze jej cech). Tak wiec, prawdopodobieństwo, że próbka o parametrach x_1, x_2, x_3,...,x_n należy do klasy X, dane jest wzorem

$$P(X|x_1, x_2, x_3, ..., x_n) = \frac{P(x_1, x_2, x_3, ..., x_n|X)P(X)}{P(x_1, x_2, x_3, ..., x_n)}$$

Słowo "naiwny"w nazwie tego klasyfikatora pochodzi od faktu, że zakłada on, że zmienne (cechy) x_1, x_2, x_3,...,x_n sa od siebie niezależne. Wtedy, wzór upraszcza sie do postaci.

$$P(X|x_1, x_2, x_3, ..., x_n) = \frac{P(x_1|X)P(x_2|X)\dot{:}..\dot{P}(x_3|X)P(X)}{P(x_1)P(x_2)\dot{:}..\dot{P}(x_3)}$$

Klasyfikacja nastepuje po porównaniu prawdopodobieństw należenia probki do danej klasy. Przewidywana klasa jest ta, której prawdopodobieństwo dla tej próbki jest najwieksze. Zatem, jako że mianownik jest taki sam dla każdego prawdopodobieństwa dla jednej próbki, klasyfikacja próbki odbywa sie na podstawie wzoru:

$$\tilde{X} = max_X(P(x_1|X)P(x_2|X)\dot{:}..\dot{P}(x_3|X)P(X))$$

2.2 Traning modelu i dane ciagłe

Proces trenowania modelu polega na znalezieniu odpowiednich wartości prawdopodobieństw na podstawie danych treninowych.

O ile znalezienie prawdopodobieństw danych dysktretnych jest bardzo intuicyjne (prawdopodobieństwo klasyczne), to jeśli chodzi o ciagłe, w moim klasyfikatorze zastosowałam dla nich rozkład normalny. Na podstawie danych treningowych, dla każdej cechy z próbki program oblicza średnia i odchylenie standardowe w każdej z możliwych klas i na ich podstawie określa prawdopodobieństwa $P(x_i|X)$.

3 Hiperparametry

Model nie ma żadnych hiperparametrów.

4 Badanie jakości działania klasyfikatora

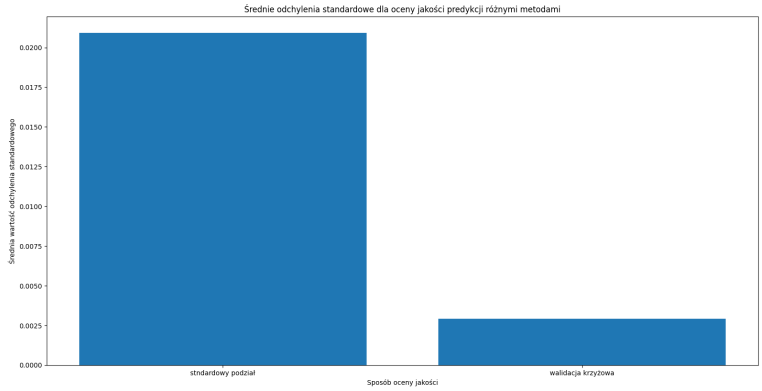
4.1 Ostateczny wynik

Poprawność przewidywań mojego klasyfikatora zbadałam dzielac zbiór danych na zbiór treningowy, walidacyjny i testowy. Dla zbioru testowego, model osiagnał skuteczność około 96.49%.

Rysunek 1: Skuteczność klasyfikatora na zbiorze testowym

4.2 Walidacja krzyżowa

Zbadałam różnice w ocenie klasyfikatora za pomoca k-krotnej walidacji krzyżowej w porównaniu ze zwykłym, stałym podziałem danych. Dla obydwu tych metod zbadałam odchylenie standardowe wyników. Uruchamiałam walidacje dla stałego podziału z różnymi rozłożeniami danych w podziale, a dla walidacji krzyżowej - z inna krotnością. Rezultat widać na wykresie poniżej.



Rysunek 2: Odchylenie standardowe walidacji krzyżowej i standardowego podziału

Jak widać, dla sztywnego podziału odchylenie standardowe pomiarów okazało sie dużo wieksze, co pokazuje wieksza niepewność wyników otrzymywanym właśnie ta metoda. Jakość zmierzona na zbiorze walidacyjnym może wiec zależeć od tego, jak dane zostana podzielone.

W k-krotnej walidacji krzyżowej, zbiór danych jest dzielony na k fragmentów. Modele trenuje sie na k-1 fragmentach a jakość sie mierzy na pozostałym. Ostateczny wynik jest uśredniony z k iteracji, gdzie kolejne z k fragmentów staja sie nowym zbiorem walidacyjnym. Dzieki temu, każda próbka w zbiorze bedzie oceniana, co prowadzi do wiekszej stabilności wyników.