

WSI - ćwiczenie 7.

Modele bayesowskie

20 maja 2025

1 Sprawy organizacyjne

1. Ćwiczenie realizowane jest samodzielnie.
2. Ćwiczenie wykonywane jest w języku Python.
3. Ćwiczenie powinno zostać oddane najpóźniej na 13. lub 14. zajęciach. W ramach oddania ćwiczenia należy zademonstrować prowadzącemu działanie kodu oraz wysłać kod z raportem (archiwum zip lub link do repozytorium) na adres mailowy prowadzącego.
4. Rozwiązanie ćwiczenia powinno być zgodne z szablonem z repozytorium WSI Template.
5. Dokumentacja powinna być w postaci pliku .pdf, .html albo być częścią notebooka jupyterowego. Powinna zawierać opis eksperymentów, uzyskane wyniki wraz z komentarzem oraz wnioski.
6. Na ocenę wpływa poprawność oraz jakość kodu i dokumentacja.
7. Można korzystać z pakietów do obliczeń numerycznych, takich jak *numpy*.
8. Można skorzystać z pakietu *scikit-learn* w celu załadowania zbioru danych oraz wykonania walidacji krzyżowej.
9. Implementacja powinna być ogólna, w szczególności powinna być możliwa do zastosowania do dowolnego zbioru danych o ciągłych wartościach atrybutów.

2 Ćwiczenie

Celem ćwiczenia jest implementacja naiwnego klasyfikatora Bayesa. *Uwaga: dane w zbiorze są ciągłe, należy to uwzględnić zakładając ciągły rozkład, np. normalny.*

Następnie należy wykorzystać zaimplementowany algorytm do stworzenia i zbadania jakości klasyfikatorów dla zbioru danych Spambase. *Uwaga: ze zbioru danych należy wydzielić zbiór testowy, pamiętając o rozkładzie klas. Na tym zbiorze powinna być badana jakość docelowego klasyfikatora.*

Jakość modelu należy sprawdzić:

- przez różne losowe podziały danych na zbiór trenujący i walidacyjny,
- przez k-krotną walidację krzyżową.

Należy porównać obie metody pomiaru, w szczególności średnie i odchylenia standardowe metryk mierzących jakość modelu.

Jakość ostatecznego modelu, wytrenowanego na całości danych nietestowych, należy zbadać na zbiorze testowym.

Zadanie dla chętnych

Przygotuj klasyfikator oparty na estymacji pełnej macierzy kowariancji. Porównaj jakość klasyfikatora z klasyfikatorem przygotowanym w ramach ćwiczenia.

3 Uwagi

1. Zbiór danych Spambase zawiera plik *spambase.data*, jego format jest zgodny z formatem *csv*.
2. Wszystkie wyniki wynikające z kilkukrotnego powtarzania eksperymentów powinny być opisane odpowiednimi statystykami (średnia, odchylenie standardowe, min, max).