WSI - Sieci bayesowskie

Jan Szymczak

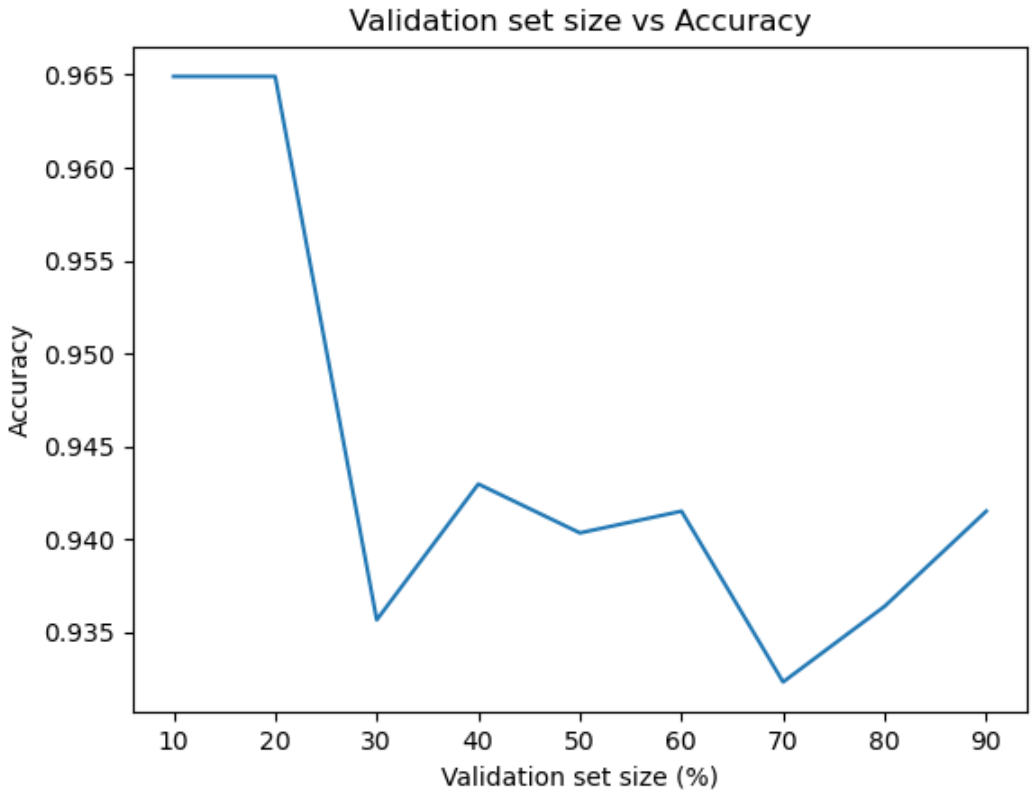
Czerwiec 2024

1 Opis problemu

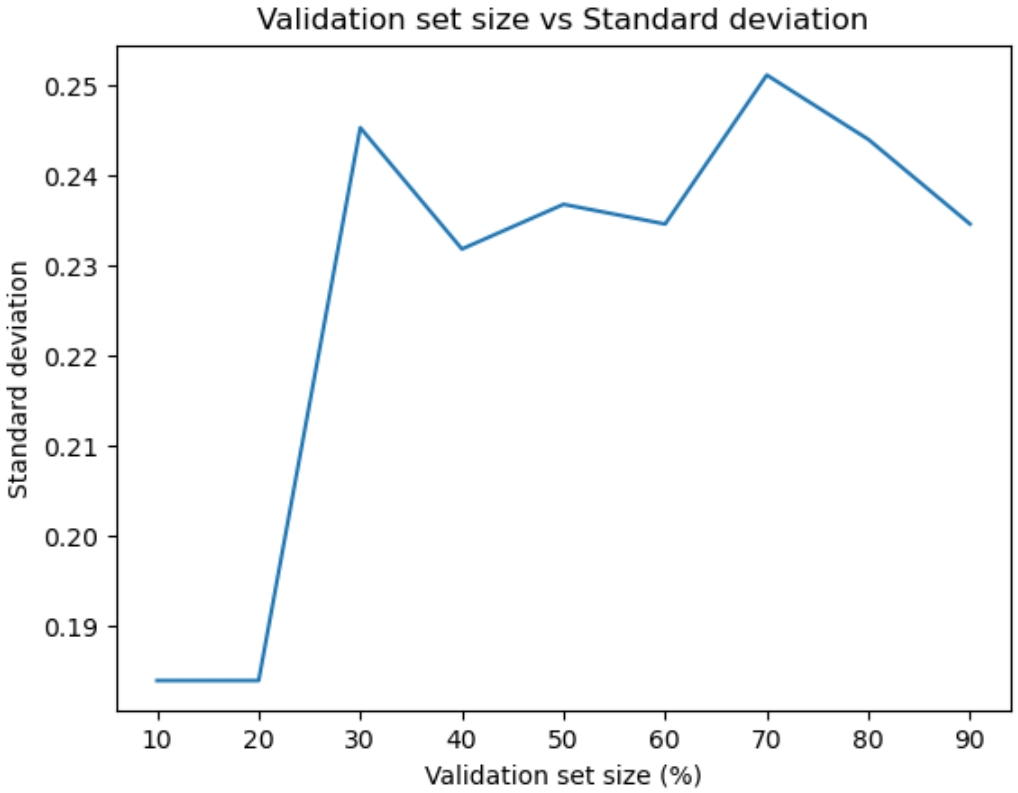
Zadaniem było napisanie implementacji naiwnego klasyfikatora Bayes'a. Następnie algorytm testowany był na danych z biblioteki scikit-learn: `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer`. Dane w zbiorze są ciągłe, dlatego wykorzystano ciągły rozkład normalny. Dane zostały także podzielone na zbiór trenujący, walidacyjny i testujący. W celu przetestowania jakości działania, przeprowadzone zostały testy na różnych proporcjach podziału zbiorów. Ponadto sprawdzono także działanie w przypadku walidacji krzyżowej.

2 Wyniki

Wyniki testowania proporcji podziałów:

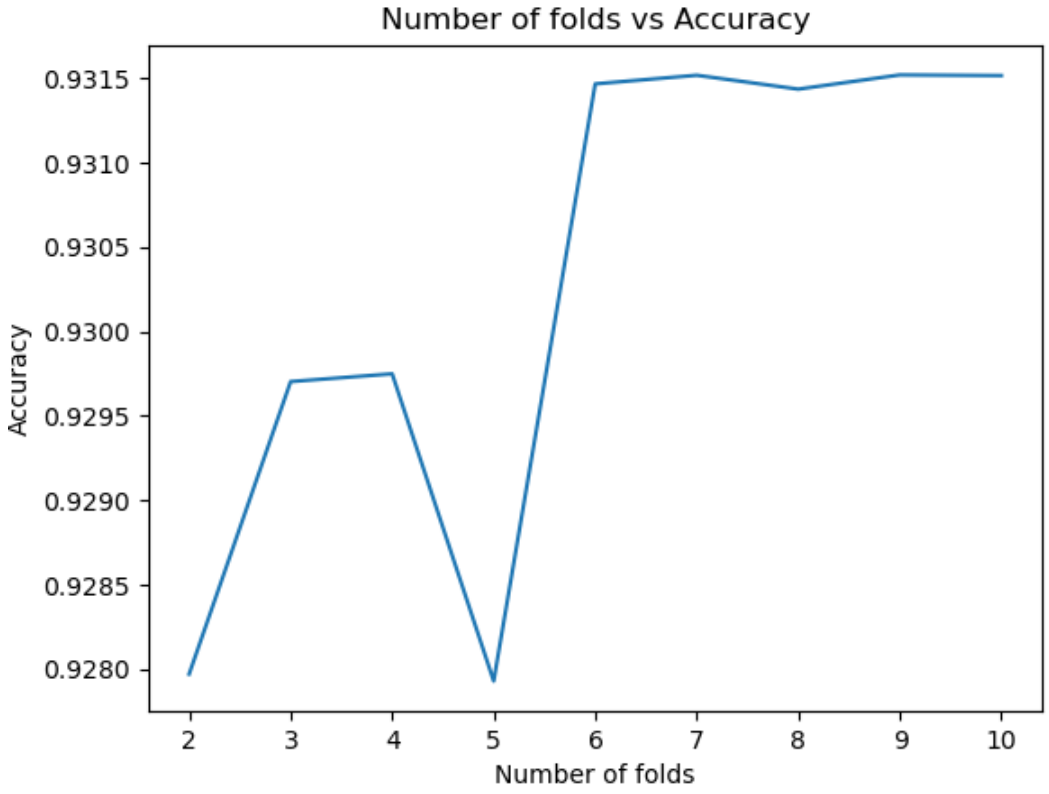


Rysunek 1: Dokładność dopasowania względem wielkości zbioru walidacyjnego

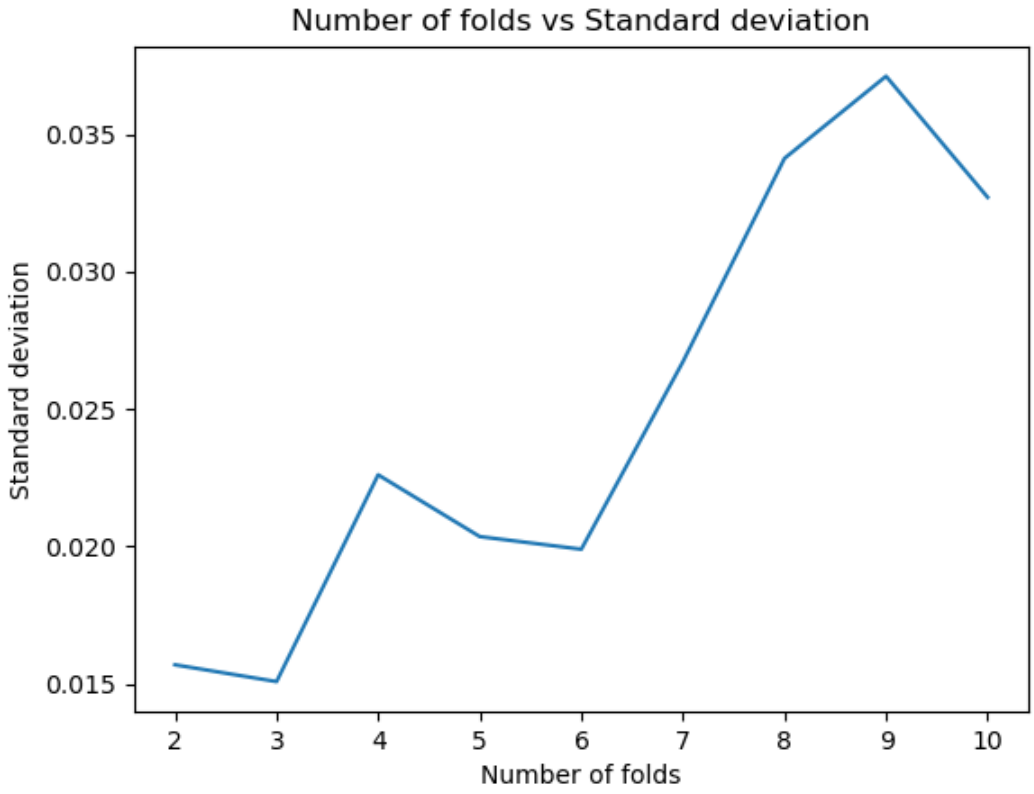


Rysunek 2: Odchylenie standardowe względem wielkości zbioru walidacyjnego

Najlepsze wyniki daje podział, gdzie zbiór walidacyjny stanowi 10% lub 20% całego zbioru danych. Dla wyższych proporcji dokładność spada, a odchylenie standardowe wzrasta, więc wyniki ulegają pogorszeniu. Wielkość zbioru uczącego jest wtedy po prostu zbyt mała, aby algorytm był w stanie nauczyć się dokładnie. W ostatecznym teście zdecydowano się na podział 80:20, ponieważ zbiór walidacyjny był jescze dzielony na pół, gdzie drugą połowę stanowił zbiór testujący.
Wyniki walidacji krzyżowej:



Rysunek 3: Dokładność dopasowania względem ilości podzbiorów, na które dzielone były dane



Rysunek 4: Odchylenie standardowe względem ilości podzbiorów, na które dzielone były dane

Najlepsze wyniki pod względem dokładności dają podziały na 6 lub więcej podzbiorów. Z drugiej strony, wraz ze wzrostem ilości podzbiorów rośnie także odchylenie standardowe. W ostatecznym teście zdecydowano się na podział na 10 podzbiorów.

2.1 Ostateczne wyniki

Ostateczne wyniki przy podziale zbiorów w proporcji 8:1:1 (trenujący, walidacyjny, testowy) oraz walidacji krzyżowej z podziałem na 10 podzbiorów:

- Dokładność na zbiorze walidacyjnym: 92.98%
- Dokładność na zbiorze testującym: 100% (z pewnością nie było to aż 100%, jednak accuracy_score() z pakietu scikit-learn stosuje pewne zaokrąglenie)
- Dokładność w przypadku stosowania walidacji krzyżowej: 93.15%

Uzyskane wyniki są świetne, otrzymane dokładności bardzo wysokie, co świadczy o poprawnym działaniu algorytmu.