MPUM mini projekt 2

Tymoteusz Siemieniuk

Kwiecień 2024

1 Przygotowanie danych

Dane zostały podzielone w następujący sposób:

- 67% - dane treningowe
- 33% - dane testowe

Zmienna objaśniana, która przyjmowała wartości 4 i 2 została przeskalowana tak, aby przyjmowała wartości 1 i 0. Nie wykonywałem żandej optymalizacji hiperparametrów, więc zbiór walidacyjny nie był potrzebny.

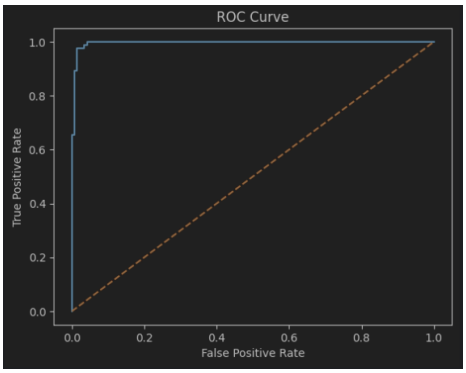
2 Wykorzystywane modele

Używam dwóch modeli: regresji logistycznej oraz naiwnego klasyfikatora bayesowskiego. Zostały one zaimplementowane w najprostszy sposób.

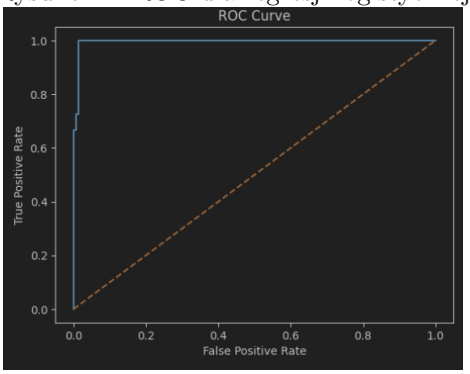
Regresja logistyczna maksymalizuje likelihood, minimalizując ujemny log-likelihood spadkiem gradientu. Otrzymywałem szybkie zbieżności i dobre wyniki dla szybkości uczenia wybranej za pierwszym razem (0.01) więc nie dodawałem ani regularyzacji, ani funkcji bazowych. Nie dodałem również early stopping ani żadnej optymalizacji spadku po gradiencie. Nie było to moim zdaniem potrzebne, gdyż model szybko osiągał dobre wyniki.

W naiwnym klasyfikatorze bayesowskim użyłem wygładzenia Laplace'a z notatek z wykładu. Klasyfikator ten napisany jest w sposób, który umożliwia mu uczenie się jedynie danych, gdzie zmienna objaśniana ma jedynie dwie wartości - 0 i 1, choć można to łatwo rozszerzyć na więcej klas.

3 Wyniki dla obu modeli



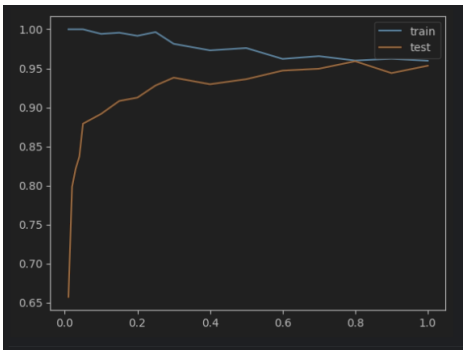
Rysunek 1: ROC dla regresji logistycznej



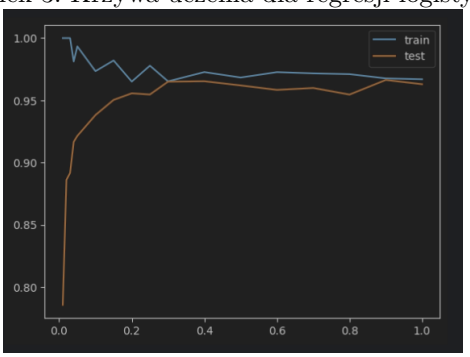
Rysunek 2: ROC dla estymatora bayesowskiego

Regresja logistyczna uzyskała pole pod krzywą ROC równe 0.9941 oraz miarę F1 równą 0.9620. Przy klasyfikatorze bayesowskim natomiast te liczby wynosiły odpowiednio 0.9944 oraz 0.9625.

4 Krzywe uczenia



Rysunek 3: Krzywa uczenia dla regresji logistycznej



Rysunek 4: Krzywa uczenia dla estymatora bayesowskiego

Na tych wykresach na osi y odkładane są miary F1 modelu wyuczonego na ułamku danych odłożonego na osi x. Wyniki zostały uśrednione na 10 przebiegach.

5 Wnioski

Krzywe ROC dla obu modeli wyglądają świetnie. Model bayesowski otrzymuje lepsze wyniki F1, co widać dobrze porównując wykresy 3 i 4. Nie wiem, czy dane które dostaliśmy nadają się dobrze na porównywanie modeli generatywnych do dyskryminatywnych, gdyż oba modele zaklepane niemalże na ślepo bez żadnych optymalizacji uzyskują dobre wyniki. Można jednak zgodzić się z jednym z wniosków postulowanych w artykule Andrew Nga i Michaela Jordana - patrząc na krzywe uczenia, rzeczywiście model bayesowski szybciej zbiega do swojego szczytowego wyniku.

Jeśli chcielibyśmy wystawić któryś z tych modeli do faktycznego użytku w diagnozie nowotworów, możnaby nie dobierać cutoffów i po prostu zwrócić praw-

dopodobieństwa - zostawić je do interpretacji dla lekarzy. Natomiast jeśli musielibyśmy dobrać jakiś kontretny cutoff, z oczywistych względów należałoby to zrobić zmniejszając FNR kosztem mniejszego TPR - dokładna liczba zależałaby już od eksperckiej opinii, nie chcę zgadywać.