Klasyfikacja nowotworów na podstawie mutacji somatycznych i ekspresji RNA

Maria Prus-Głowacka
Ćwiczenie 5 z Biologii systemów

Czerwiec 2025

1 Wstęp

Celem projektu było zbudowanie modelu klasyfikującego nowotwory na podstawie danych wieloomicznych - mutacji somatycznych oraz ekspresji RNA. Geny brane pod uwagę pochodziły z lisy Vogelstein et al. z bazy OncoKB. Dane mutacji somatycznych zostały pobrane z tabeli *masked_somatic_mutation_hg38* z BigQuery, a dane ekspresji RNA z tabeli *RNAseq_hg38_gdc_current*.

2 Metody

2.1 Przetwarzanie danych

Z genomu referencyjnego zostały wyciągnięte motywy mutacyjne. Następnie mutacje zostały znormalizowane względem pirymidyn. Następnie genomy zostały zbinowane co 1 Mb. RNA_seq zostało znormalizowane. Zbiór danych został podzielony na treningowy i testowy względem proporcji 80/20.

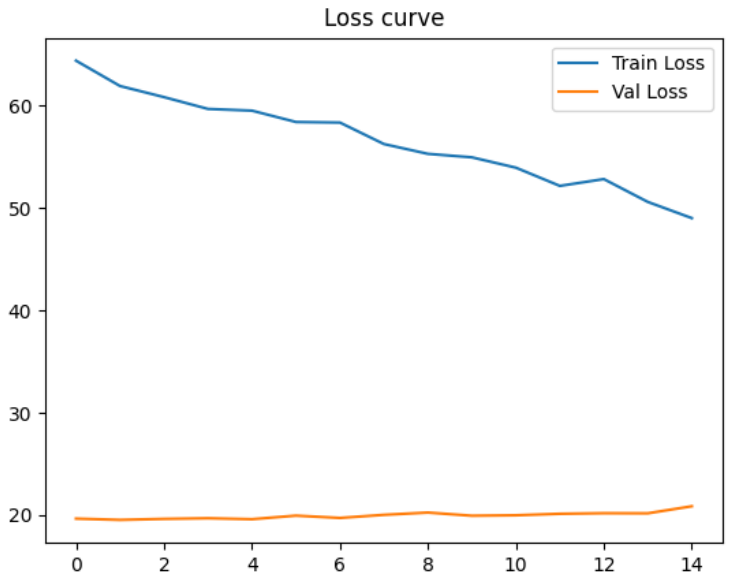
2.2 Model

Do klasyfikacji wykorzystano głęboką sięć neuronową zaimplementowaną w bibliotece PyTorch. Architektura składała się z trzech warstw liniowych połączonych z mechanizmami normalizacji, aktywacji nieliniowej oraz regularyzacji. Pierwsza warstwa mapowała dane wejściowe na przestrzeń o wymiarze 256. Następnie zastosowana zostaje warstwa normalizacji wsadowej (Batch Normalization), która normalizuje wyjście w odniesieniu do bieżącej partii danych, zmniejszając problem tzw. internal covariate shift. Dzięki temu stabilizuje i przyspiesza proces uczenia się. Następnie dane przechodzą przez funkcję reaktywacji ReLU oraz warstwę Dropout z prawdopodobieństwem $p = 0.3$, co wprowadza losową deaktywację neuronów w trakcie treningu, pełniąc funkcję regularyzacji i zapobiegając przeuczeniu modelu. Kolejna warstwa redukuje wymiar do 128 a dane analogicznie przechodzą przez normalizację, reaktywację oraz regularyzację. Warstwa wyjściowa mapuje dane z przestrzeni 128-wymiarowej

do przestrzeni klasyfikacji o wymiarze liczby klas. Ostateczna klasyfikacja wykonywana jest przez funkcję Cross Entropy Loss. Trening trwał przez 15 epok.

3 Wyniki

Najlepsza otrzymana dokładność wyniosła około 41%. To oznacza, że model znacząco przewyższył losową dokładność, która przy 24 klasach wynosi około 1,4%. Rysunek 1 przedstawia wykres wizualizujący krzywe *loss/accuracy* dla zbioru treningowego i walidacyjnego. Widać, że strata treningowa stopniowo maleje z poziomu około 60% do 50% a strata walidacyjna trzyma się na poziomie około 20%.



Rysunek 1: Caption