**PREDICTING SIGNAL PEPTIDES**

Sheng Han, Feride Eren

**Abstract**

Prediction on the existence of signal peptides requires computational approach because they are not responsive to the current crystallization methods. 4 different prediction algorithms are used and compared in this project which are Bernoulli NB, Multinomial NB, Logistic Regression and Linear Support Vector Machine. A stretch of hydrophobic amino acids is both present in signal peptides and transmembrane helices which makes it challenging for signal peptide and transmembrane helix predictors to correctly distinguish the stretches of hydrophobic residues near the N-terminal methionine of a protein sequence. The results for each predicting model revealed logistic regression with n-gram of 3 to 4. Our final model achieved 84% accuracy on the testing set.

**Author Summary**

Not sure about this part

**Introduction**

Signal peptides are extra peptide extension that contains 16–30 amino acids added to the N-terminus of the protein. They function by helping the transport mechanism within the cell to bring it to its specific destination which protein is delivered within the cell. It is removed while the protein is translocated across the endoplasmic reticulum membrane.

One of the most important topics to investigate in bioinformatics is structure and function of membrane proteins. Especially transmembrane proteins control very important function in organism (Yu and Zhang, 2013). To put simply, flow of information and substances in and out of cell is controlled by transmembrane proteins which are also important drug targets (Reynolds et al., 2008). 70% percent of the drug target that are known or being researched are transmembrane proteins (Yildirim et al., 2017)

Confusions at predicting transmembrane protein can occur especially at signal peptides’ strongly hydrophobic and not part of the mature protein region is misclassified as a membrane-spanning portion of a transmembrane protein (Reynolds et al., 2008). These kinds of inaccurate analysis also happen when a transmembrane protein with a membrane-spanning segment near the N-terminus is often misclassified as having a signal peptide.

In this project, it is aimed to train and test a signal peptide classifier based on the data provided and two proteome sets chosen, Drosophila melanogaster and Mus musculus. Note that this project does not aim at predicting the exact locations of signal peptides but only their existence.

Prediction models used in this project are Bernoulli Naïve Bayes (NB), Multinomial NB, Logistic Regression and Linear Support Vector Machine (SVM).

**Methods**

Two probabilistic models and two linear models are considered in this project. The two probabilistic models are generative models based on Naïve Bayes assumption, where it is assumed that each of the peptide is considered in-depend from others. Although it is known to be false, this model is commonly used in Machine Learning and have been proven useful.

The two linear models – Logistic regression and SVM are also two commonly used linear classifier. Both model attempts to separate the training data by constructing a decision boundary around the known samples. For SVM, different kernel will be used, which transform input data into higher dimensions thus made linear separate possible in some cases.

## Training Data

Each of the 4 models are trained with provided labelled data. With a 0.2 testing set and 0.8 training set split. Fasta format input data are divided and then vectorised to create sequence embedding. A simple Count Vectoriser is used to transform raw text into matrices according to the number of occurrence for each of the unique amino acid in any given protein sequence.

Since there are total of 21 amino acids, this results an input vector of N by 21. Where N is the total training sample.

To improve the accuracy and also the ratio between number of features and samples, N-gram is used to combine adjacent amino acid into groups of k members. Where k is the number of amino acid in a group. This k is a hyper-parameter that requires tuning during training.

Transmembrane (TM) protein dataset and non-TM dataset are trained separately at first, then combined for prediction. It is expected the TM protein dataset will perform poorly compared with non-TM dataset.

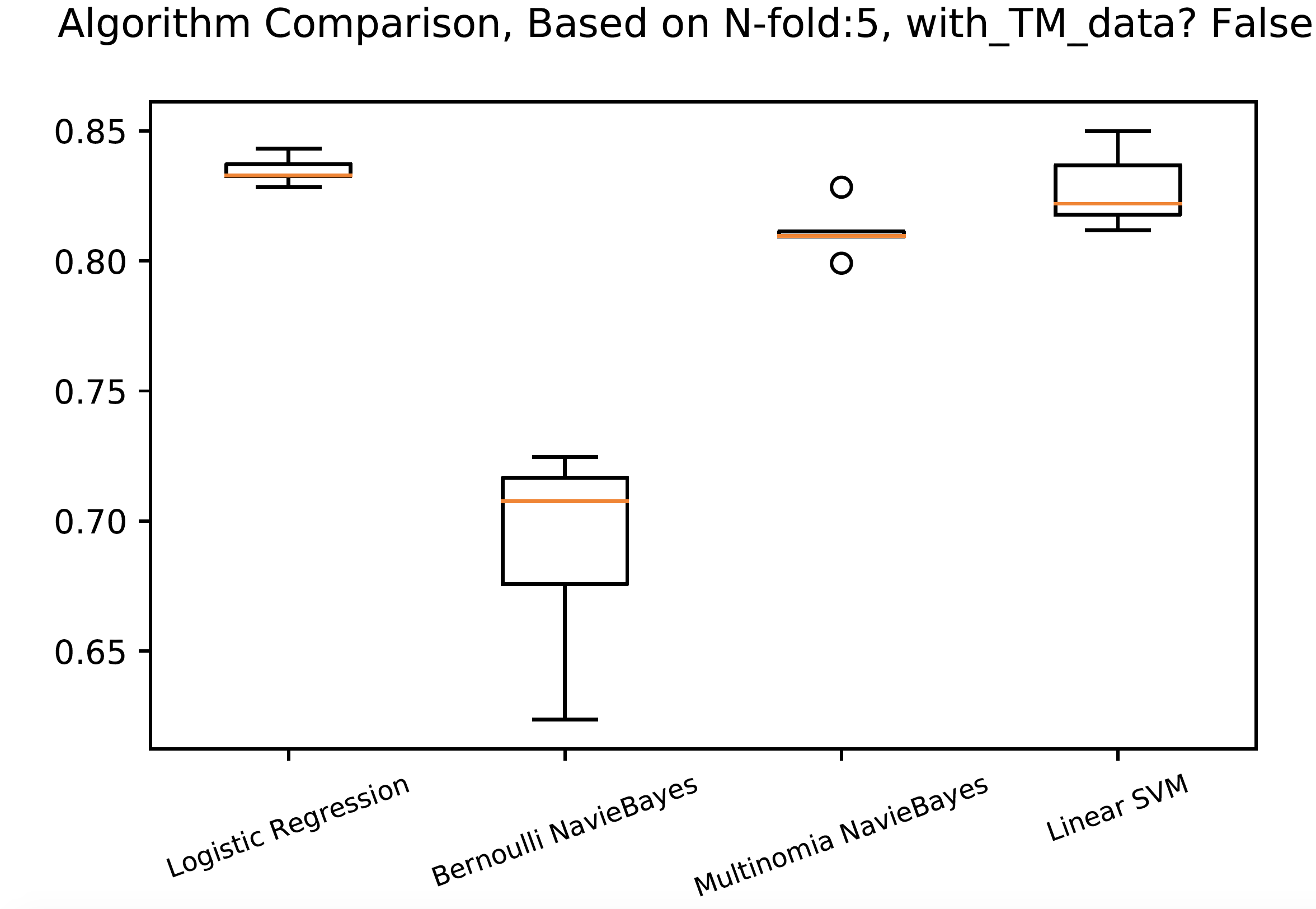
## Model Evaluation

There are two model evaluation methods used. The first is called performance test, where the accuracy of each model is compared based on n-fold cross-validation. This provide an overview of the true performance in terms of predicting accuracy.

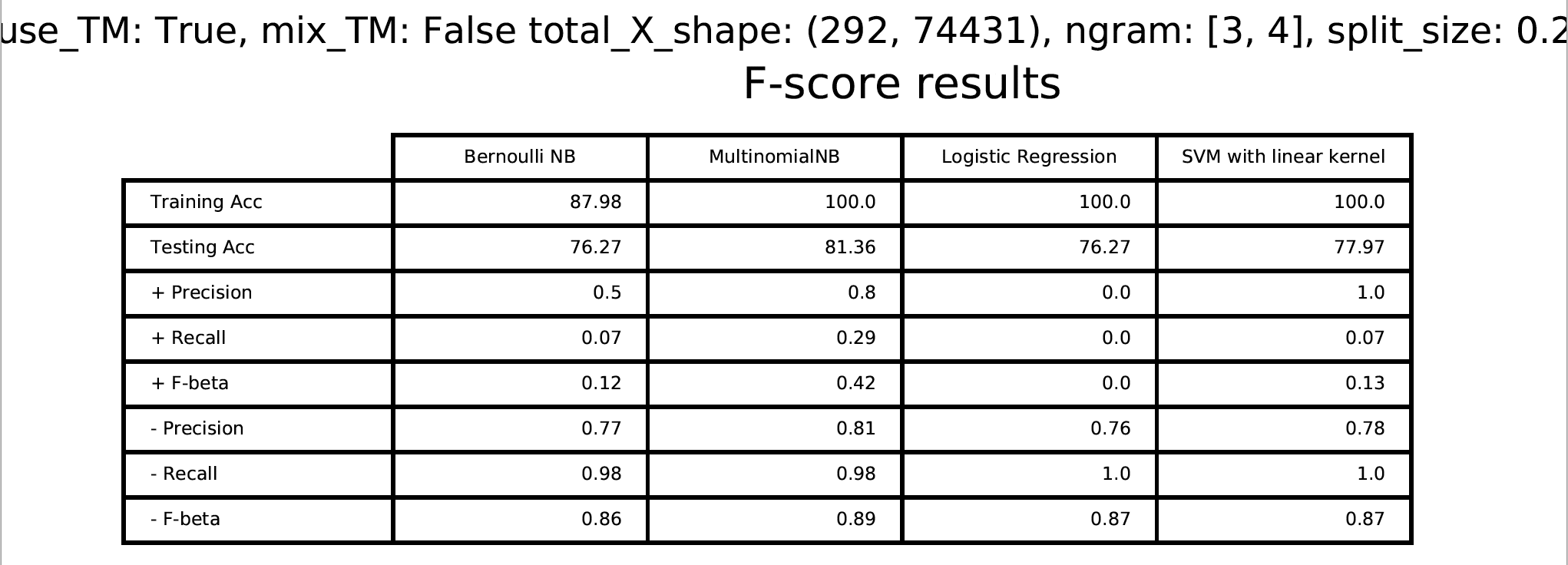
In order to gain additional insights, precision and recalls must be used. Those measures provide a detailed performance break down in terms of the false negative, false positive, true positive and true negative.

**Results**

The hyper-parameter turning revealed the best n-gram is when k = 3 and 4. The best kernel for SVM is a linear kernel. All results are saved under /results/ folder. SVM, Multinomial NB and logistic regression have similar performance in term of raw accuracy. Bernoulli NB performed the worst among the 4.



The performance on TM-data is very poor when compared with non\_TM data. This can be seen from the f-score test below.



All models performed poorly on TM-dataset, which may cause by the smaller dataset (292) when compared with non-TM dataset (2362). From the table above, positive f-score are poor comparing with negative score. This indicate all models failed to classify positive samples correctly.

**Discussion**

**References**

Yıldırım, M., Goh, K., Cusick, M., Barabási, A. and Vidal, M. (2007). Drug—target network. *Nature Biotechnology*, 25(10), pp.1119-1126.

Yu, B. and Zhang, Y. (2013). A Simple Method for Predicting Transmembrane Proteins Based on Wavelet Transform. *International Journal of Biological Sciences*, 9(1), pp.22-33.