BIO310 Introduction to Bioinformatics
HW4
Spring 2019

May 4, 2020

Instructions:

- This assignment is HW4. Submit your answers as an `ipynb`. Include markdown where necessary.
- We will perform a tutorial related to this in the recit. That might be helpful for completing this homework.

Coding Potential Prediction

Ribo-nucleic acid (RNA) molecules perform many different functions in biological systems.

Some RNA molecules can be translated into protein in order to perform a specific function in the cell, in which case the RNA is said to *code* for a protein and is called a *coding RNA*. Some RNAs are not translated into protein and perform functions in the cell in their native RNA form, such RNA are known as *non-coding RNA*.

In this exercise, we'll build a classifier that can distinguish between coding and non-coding RNA based on features generated from the nucleotide sequences. The features we use here are taken from [1].

For example, given a sequence like `ATGCTTGCA...` one set of features could be the counts of how many As, Ts, Gs and Cs there are in this sequence. Another set of features could be to count the number of dimers, e.g. how many `AA`s, `AT`s, `AG`s, `AC`s and so on there are in this sequence. If you're curious about other types of features used, refer to the paper.
In this assignment, you will implement a random forest classifier. This classifier will be trained to distinguish between coding and non-coding RNAs based on the features calculated from the nucleotide sequences (available in `coding_potential.csv`).

Follow the following steps to complete this assignment:

- **Import data**. Read the .csv file provided. The `label` column is what we are trying to predict (`target, y`) using the rest of the columns, (`features, X`). If the label is `1` for a given row, it means that the features in that row were calculated from a *coding* RNA. Similarly, if the label is `0`, it means that the features in that row were calculated from a *non-coding* RNA.
- **Split the data into train and test sets.**
- **Train a random forest model.** Make sure you use only the training data in this step.
- **Evaluate the model**
  - Make predictions on the *training* data using the trained model. Generate a confusion matrix and classification report.
  - Make predictions on the *test* data using the trained model. Generate a confusion matrix and classification report.

Answer the following questions:

1. How many rows and columns does the `coding_potential.csv` have?
2. When you split the data into segments, how many samples (rows) are in the training and test data?
3. Report the confusion matrices and classification reports for the training and test data.
4. Explain what the confusion matrix describes.
5. What are precision, recall and F1-score?
6. Does the classifier perform better when predicting the training data or the test data targets?
7. Why is there a difference between the training and test prediction accuracies?

References

[1] Xiaoxue Tong and Shiyong Liu. Cppred: coding potential prediction based on the global description of rna sequence. *Nucleic acids research*, 47(8):e43–e43, 2019.