GS 373 Homework 8

Due May 31st before 1:30 PM on Canvas

- (100 points): 4 bioinformatics questions (80 points), 1 programming assignment (20 points).
- Submit answers to the bioinformatics questions in a Microsoft Word document or PDF via Canvas. Your answers do not need to contain the text of the questions, but they need to be clearly labeled (e.g., 1a., 3b., etc.)
- Submit the programming assignment as a separate .py file onto Canvas. The script should be able to be directly run by Python.

Bioinformatics Questions (80 points)

1. (25 points) Suppose you want to **predict whether a single amino acid variant in a protein is damaging or neutral from two binary features: whether the site is conserved and whether the site is solvent exposed**. Your training data consists of a database of **2000 variants of known effect (1000 damaging, 1000 neutral)** that are described in the table below. (According to the table, 46 of the damaging variants are conserved and solvent un-exposed)

# of damaging variants, # of neutral variants

Feature	Solvent exposed	Solvent un-exposed
Conserved	754, 101	46, 99
Not conserved	162, 697	38, 103

   - (5 points) In this data, **which feature is more informative** (results in purer separation) about the effect of the variant?
   - (10 points) **Construct a binary decision tree** based on these features, using the greedy algorithm shown in class. **How many variants in the training data would this tree predict to be damaging? How many variants would be predicted to be neutral?**
   - (5 points) What is the **accuracy** of the decision tree on its own training data?
   - (5 points) What is the **true positive rate**? What is the **true negative rate**?

2. (15 points) Suppose you are doing a metabolic engineering project where you want to find ways to increase the activity of a particular pathway in *Saccharomyces cerevisiae* yeast. **You decide to build a machine learning model to predict whether knocking out different genes will affect the expression of your gene of interest.** In other words, every data point is a single-gene knockout strain, and the outcome to predict is whether or not that strain showed higher flux through your target pathway. **Propose 3 features or classes of features that could be potentially informative for this model.** The features can be derived from any experimental assay or prior knowledge of the knockout genes and their relationship with the target pathway. **Explain how each one could be informative.**

3. (20 points, 5 points each) **State which type of machine learning algorithm** (either classification, regression, or clustering) would be most appropriate for each analysis below and **explain how you would use it by specifying what the objects and features would be**.
    - You have single-cell RNA-seq data for the cells in a tissue sample and want to know how many cell types exist in the tissue.
    - You want to infer the DNA sequence associated with current flow data from Nanopore sequencing given the known current flow associated with different nucleotides.
    - You want to understand how the expression of multiple genes affects the expression of another gene, given multiple expression measurements from many cell types.
    - You want to identify metabolic biomarkers of insulin resistance, based on a serum metabolomics assay applied to samples from pre-diabetic and healthy individuals.

4. (20 points) Describe what metric you would choose to optimize and in which direction (higher true positive but higher false positive or lower true positive and lower false positive), if you were using machine learning for each of the following classification problems. Explain why.
    a. You are designing a classifier for early detection of Ebola virus.
    b. You are designing a classifier for early detection of a slow-growing cancer.

Programming (20 points)

Write a program to analyze prediction results of a binary classifier. Your program should read in a table with three columns, separated by spaces: sample names, their true classification values (0 or 1), and predicted classification values from some machine learning model (again 0 or 1). It should perform the following steps:

1) Read in the input file (specified from `sys.argv`) and store the data in an appropriate structure or structures. Remember to convert the classification values from strings into numeric or Boolean data types.
2) Print the names of all samples that were classified incorrectly.
3) Calculate the number of true positives, true negatives, false positives, and false negatives in the input dataset.
4) Define 3 functions to calculate accuracy, TPR, and TNR (remember to place these at the beginning of your program). You can decide how to format the arguments for these functions. Apply each function to the input data and print the results.
5) Write a 2x2 contingency table based on the data to an output file (filename provided from `sys.argv`), again with each value separated by spaces. The rows should represent true positives and negatives (in that order), and the columns should represent positive and negative model predictions (in that order).

Example contents of `input_file.txt`:

```
sample1 0 0
sample2 1 0
sample3 1 1
sample4 1 0
```

Example contents of `output_contingency.txt`:

```
1 2
0 1
```

Example usage and output printed to terminal:

```
python homework8_skh.py input_file.txt output_contingency.txt
```

```
Incorrectly classified samples:
sample2
sample4
Accuracy: 50%
TPR: 33%
TNR: 100%
```