Machine Learning Project Description

Date: November 23, 2023

1 Project 1: Breast Cancer Driver Gene Prediction

Breast cancer is one of the most frequently diagnosed cancers in women, in which the cells in the breast grow uncontrollably. This growth is determined by mutations in certain genes. A mutation that creates a fundamental growth advantage is called a driver mutation, and the associated gene is called a driver gene. Passenger mutations, on the other hand, do not necessarily determine the development of cancer. Uncovering driver genes will not only help improve our understanding of biological mechanism underlying cancer development and progression, but also develop new treatments. To accelerate this process, machine learning models are needed to predict potential driver genes from a large set of candidate genes.

The dataset **P1** is a binary and heavily unbalanced dataset which contains 2435 examples (genes). Each gene is represented by 12 mutation features.

Your task in this project is to build classification models to classify each gene into driver class (label 1) or passenger class (label 0) using different machine learning techniques these have been studied in the course. The experimental setup needs to cover following points.

- Using 5-fold nested cross validation
- Considering to use sampling techniques
- Employing AUPRC as the metric to measure the performances of models since the dataset is unbalanced
- Performing hyper-parameter optimization for choosing optimal hyper-parameter tuple for each model

In the end, you are asked to make a presentation (maximum in 20 minutes) which includes:

1. Introduction to the problem, highlighting the importance of the task.
2. An overview of the methods used to build your models
3. Experimental setup
4. Results which show the performances and comparison of different models.
5. Conclusion

2 Project 2: RNA-binding Protein Site Prediction

RNA-binding proteins (RBPs) regulate many vital steps in the RNA life cycle, such as splicing, transport, stability, and translation. Numerous RBPs have been implicated in diseases like cancer, neurodegeneration, and genetic disorders, urging the need to speed up their functional characterization and shed light on their complex cellular interplay.

An important step to understand RBP function is to identify the precise RBP binding locations on regulated RNAs. CLIP-seq and its popular modifications such as PAR-CLIP, iCLIP, and eCLIP has become the state-of-the-art technique to experimentally determine transcriptome-wide binding sites of RBPs. Despite their ability to accurately identify RBP binding sites, their high cost is of a big concern. Therefore, computational methods have been proposed to help speeding up this determination process.

The dataset **P2** consists of 3101 positives and 3101 negative RNA-binding protein sites regarding PUM2 protein.

Your task in this project is to build binary classification models to classify each site into the positive or negative class using different machine learning techniques these have been studied in the course. The experimental setup needs to cover following points.

- Using 5-fold nested cross validation
- Employing different appropriate evaluation metrics to measure the performances of models
- Performing hyper-parameter optimization for choosing optimal hyper-parameter tuple for each model

In the end, you are asked to make a presentation (maximum in 20 minutes) which includes:

1. Introduction to the problem, highlighting the importance of the task.
2. An overview of the methods used to build your models
3. Experimental setup
4. Results which show the performances and comparison of different models.
5. Conclusion

3 Project 3: Stock Price Prediction

The Google Stock data (**P3**) has information from 19/Aug/2004 to 24/Mar/2022. There are five columns. The Open column tells the price at which a stock started trading when the market opened on a particular day. The Close column refers to the price of an individual stock when the stock exchange closed the market for the day. The High column depicts the highest price at which a stock traded during a period. The Low column tells the lowest price of the period. Volume is the total amount of trading activity during a period of time.

Your task in this project is to build regression models to predict the opening price using different machine learning techniques these have been studied in the course. The experimental setup needs to cover following points.

- Employing different appropriate evaluation metrics to measure the performances of models
- Performing hyper-parameter optimization for choosing optimal hyper-parameter tuple for each model

In the end, you are asked to make a presentation (maximum in 20 minutes) which includes:

1. Introduction to the problem, highlighting the importance of the task.
2. An overview of the methods used to build your models
3. Experimental setup
4. Results which show the performances and comparison of different models.
5. Conclusion