Machine Learning in Network Biology
Homework 1

Dr. Amin Emad

Gian Favero: 261157396

October 6th, 2023

Contents

1 Introduction

Transcriptomic data refers to a collection of RNA molecules, or transcripts, present in cell tissues at a reference time. Typically, these molecules are generated in a cell through the process of transcription of DNA into RNA. This data contains insights into the expression levels of genes in a cell which can be used to predict various items like disease, stimuli response, and development, among others.

Analyzing transcriptomic data is a complex task due to the large number of genes present in a cell. However, successful analysis of this data can lead to important discoveries in gene function, networks, and pathways involved in biological processes. This is especially important in the field of network biology where the goal is to understand the structure and function of biological networks.

Several classical machine learning approaches can be used when analyzing transcriptomic data. These include dimensionality reduction, clustering, and regression. All three of these particular approaches will be explored in this assignment using popular Python libraries to gain some insight into the data provided. Gene expression data was provided alongside this assignment within a file called
`gdsc_expr_postCB.csv`. The file contains rows that begin with a gene name and are taken to be features of the data, while columns represent gene lines, or instances.

2 Part 1: Dimensionality Reduction

Dimensionality reduction is a technique used to reduce the number of features in a dataset. This is done by projecting the data onto a lower dimensional space and finding a new, simpler set of features that can be used to represent the original data. Memory usage, computational efficiency, and visualization are all advantages of dimensionality reduction.

Several dimensionality reduction techniques are explored in this assignment. These include Principal Component Analysis (PCA), UMAP, and t-SNE algorithms.

2.1 Principal Component Analysis

A PCA algorithm was implemented using the `sklearn.decomposition.PCA` library in addition to peripheral libraries for data manipulation (`numpy`) and visualization (`matplotlib`).