A bunch of random studies on cancer

Adrin Jalali

September 3, 2015

Contents

Chapter 1

Introduction

Here we introduce.

Chapter 2

Background

2.1 Machine Learning

Machine learning techniques are used to extract information from data, or make some predictions about the data. This chapter briefly explains methods and techniques used in, or required to understand the proceeding chapters.

2.1.1 Feature Selection

Feature selection is the task of selecting important features to the problem at hand. It becomes particularly a hard task when the number of features in the data is of a higher magnitude compared to the number of given samples. Table 2.1 shows an example number of samples vs. number of features in a typical data. One of the challenges when dealing with such a large number of features is that if enough number of features have a probability distribution independent of the outcome, some of them might falsely seem correlated with the outcome. Another obstacle comes from the fact that our features are not independent and they function in complex networks. As a result, features should be considered in groups, which is a combinatorial and intractable problem.

Sample Data		
Sample Count	Gene Expression Data Feature Count	450K Methylation Chip Data Feature Count
500	$\approx 20,000$	$\approx 450,000$

Table 2.1: An example number of samples and features in our usual data

We have used correlation [**?**], mutual information [**?**], and $l1$-regularized methods [**?**] as techniques to select features.

2.1.2 Classification

Classification is the problem of putting data into different classes [**?**]. During the training phase, the matrix $X_{samples \times features}$ is given as the input and $y_{samples}$ as the desired output. The vector y has values from a discrete set. If the set has only two distinct values, the problem is called a binary classification.

Logistic regression [**?**], Support Vector Machines (SVM) [**?**], and decision trees [**?**] are examples of classification methods.

2.1.3 Regression

In statistics, predicting a continues output value given an input data is called regression [**?**]. Regression and classification differ in their desired output type. In regression the output is continues in contrast to classification in which the output is a discrete value.

2.1.4 Regularization

2.1.5 Empirical Risk Minimization

2.1.6 Kernel Methods

2.1.7 Gaussian Processes

2.1.8 Boosting and Ensemble Methods

2.2 K Best Shortest Paths

2.3 Lymphoma

Chapter 3

Flow Cytometry Analysis

Here we talk about flow cytometry data, how we analyze and visualize it; and how we use that analysis alongside with some machine learning tools to classify samples into cancer subtypes.

3.1 Flow Cytometry

3.2 Data Preprocessing and Challenges

3.3 High Dimensional Analysis and Visualization

3.3.1 flowType

3.3.2 RchyOptimyx

3.4 Lymphoma Diagnosis Quality Checking

Chapter 4

Adaptive Learning

Here we talk about adaptive and interpretable methods.

4.1 Challenges in Cancer Data

4.1.1 Cancer Heterogeneity

4.1.2 Batch Effects and Noise

4.2 RatBoost

4.3 Raccoon

Chapter 5

Conclusion