A bunch of random studies on cancer

Adrin Jalali

September 6, 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.

Linear regression [**?**], Gaussian processes [**?**], and kernel based regression [**?**] are some available methods here.

2.1.4 Regularization

Building a machine to predict the outcome with a good performance on the training set is easy if the number of features in the data is large enough compared to the number of samples, even if features are drawn from a random background probability distribution independent of the outcome. But the trained machine will perform poorly on the unseen test samples. This phenomenon is called overfitting. One way to prevent overfitting is to select potential features before training a selected model.

Another way to tackle the problem is to modify the models in a way that they tend to use less number of features and extract less complex patters from the data. The intuition behind regularization is that having a simple explanation for the data is probably a closer estimation to the reality than a complex one.

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

Flow cytometry is a technology that allows measurement of biomarkers inside and outside cells on a single cell basis [**?**]. The technology can also sort certain cells according to a given criterion [1].

Cell preparation in flow cytometry involves suspension of the cells in a liquid containing biomarker reagents. Reagents are marked antibodies that can be detected by the laser beams in the flow cytometer machine [**?**]. The antibodies are usually marked with a fluorescent label. Each fluorescent marker has a corresponding peak excitation and emission wavelength which can be detected using lasers available on the flow cytometry machine. The combination of markers has to be chosen such that their corresponding wavelengths have minimal overlap; otherwise they cannot be distinguished from one another.

3.2 Data Preprocessing and Challenges

Transformation and spillover compensation are the two main phases of raw flow cytometry data preprocessing.

Transformation: The measured fluorescent intensities almost exponentially correspond to the number of existing fluorescent markers on or inside the cell. Therefor a proper transformation of the raw data is essential in order to have the data in a linear space.

Spillover Compensation:

Flow cytometry data usually suffers from batch effects, and the batch effect comes from at least three different sources:

- *Reagent and Solution Batches*: The smallest variation in the solution used to stain the target cells can affect the amount of antibodies attached to target cells.

- *Laser Intensity*: The intensity of the laser used to detect stained antibodies affects values read by the instrument.

- *Compensation Matrices*: The compensation matrix is used to post-process the data and to correct for overlapping spectrum of the color of the lasers.

Compensation is necessary due to spillover. `https://www.bdbiosciences.com/documents/Compensation_Multicolor_TechBulletin.pdf`

> What are spillover and compensation? When using multiple fluorochromes in an experiment, there are many factors that can impact the accuracy and quality of the data. The most critical factor is determining which color should be matched to each antibody in the reagent panel. This is due to a very large range of intrinsic brightness among the fluorochromes commonly used, some antigens being dimly expressed while others brightly expressed, and signals from one reagent optically interfering with signals from another. These choices of color and antibody must also be made in the context of which markers might be coexpressed on the same cells. Whenever more than one marker is expressed on a single cell, the presence of the other fluorescent reagents can contribute significant optical background in proportion to their brightness. This phenomenon is called spillover.
>
> Spillover is due to the physical overlap among the emission spectra of certain commonly used fluorochromes.
>
> Spillover occurs whenever the fluorescence emission of one fluorochrome is detected in a detector designed to measure signal from another fluorochrome (Figure 1).
>
> The amount of spillover is a linear function, so the measured average signal levels can be corrected (ie, the population medians aligned) by the process called compensation.
>
> With proper compensation setup, complex data sets can then be properly visualized and analyzed if a well chosen immunofluorescent reagent panel is used. If the compensation is incorrect, interpreting the data can become extremely difficult or impossible.

Usually the data comes with its corresponding compensation matrix.

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

Bibliography

[1] Mack J Fulwyler. Electronic separation of biological cells by volume. *Science*, 150(3698):910–911, 1965.