A bunch of random studies on cancer

Adrin Jalali

September 7, 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 [8], mutual information [11], and $l1$-regularized methods [7] as techniques to select features.

2.1.2 Classification

Classification is the problem of putting data into different classes [6]. 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 [14, 2], Support Vector Machines (SVM) [13, 1], 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 [6], 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 and separate certain cells according to a given criterion [4].

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 or lamps available on the flow cytometer 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 due to interference between them.

In a flow cytometer cells flow in a liquid stream one by one, where a lamps or laser beams in conjunction with sensors measure the intensity of reflected light from the cells. These measurements can be in linear or logarithmic space [12]. The measured values depend on the light intensity projected onto cells which can be tuned by changing the voltage of the lasers or lamps. Different wavelengths correspond to different markers, but they might overlap. When the tail of the emission spectrum of a marker overlaps with the main part of the emission spectrum of another marker, it is called spillover as shown in Fig. 3.1 [10].

Compensating for spillover requires a spillover matrix (SM). $SP_{i,j}$ shows the percentage that marker i spills over marker j. The compensation matrix (CM) is then the calculated as the inverse of the spill over matrix. Let S be the true signal value, and O be the observed value. Then we have [1]:

$$CM = SM^{-1} \tag{3.1}$$

$$S = O \times CM \tag{3.2}$$

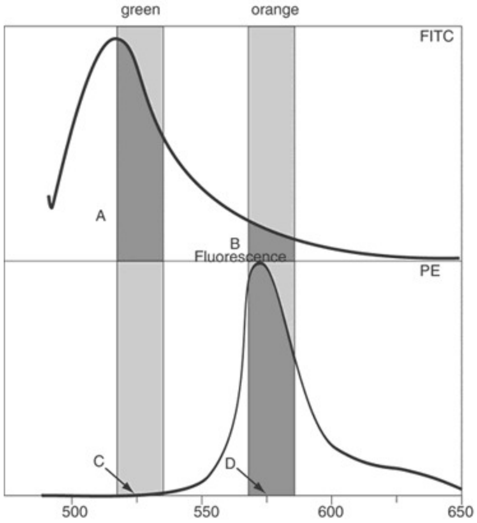

Figure 3.1: [DIRECT QUOTE] Fluorescence emission spectra for FITC and PE. The emission spectrum (the wavelengths of light generated by excitation of these molecules) is shown for an excitation at $488nm$ (the same as the argon-ion laser line). FITC emission is maximal at $\sim 515nm$; typically, a filter centered on $530nm$ is used to collect the emitted light (shaded region). The emission of is farther red, with a maximum at $\sim 575nm$; typically, a filter centered on this emission maximum is used to collect. Note that PE has some emission in the wavelength bands used to collect PE fluorescence (B); typically, the amount of light in the $575nm$ band is $\sim 15\%$ of that in the $530nm$ band (A). The PE has very little emission in the $530nm$ band (C), usually less than 2% of the emission in the $575nm$ band (D) [10].

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.

[1] http://bioinformin.net/cytometry/compensation.php

Therefor a proper transformation of the raw data is essential in order to have the data in a linear space. Logarithmic, log-linear hybrid transformation Logicle [**?**], and hyperbolic arcsine [**?**] are some commonly used transformations. Some studies have compared different transformation techniques and reported their advantages and disadvantages [3, 9].

Spillover Compensation: Compensation is done as shown in Formula 3.2 and it relies on a given compensation or spillover matrix.

In practice data are produced through time and also maybe in different labs. This means reagent batches are different, and also flow cytometry machines are not necessarily calibrated alike, which also affects compensation matrices. Therefore normalization is a crucial step to make samples comparable [5].

3.3 High Dimensional Analysis and Visualization

Manual analysis of flow cytometry data involves *gating*. Researchers use density or scatter plots of one or two selected dimensions of flow cytometry data in order to visualize and also select some areas on those plots to further investigate cells within the selected area. Visualization and further gating of those selected cells is commonly a next step to the analysis.

Manual gating of cells across several samples is a labor intensive and time consuming process. Not being able to analyze the data in its original higher dimensional space is another disadvantage of manual flow cytometry data analysis.

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] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[2] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

[3] Greg Finak, Juan-Manuel Perez, Andrew Weng, and Raphael Gottardo. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC bioinformatics*, 11(1):546, 2010.

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

[5] Florian Hahne, Alireza Hadj Khodabakhshi, Ali Bashashati, Chao-Jen Wong, Randy D Gascoyne, Andrew P Weng, Vicky Seyfert-Margolis, Katarzyna Bourcier, Adam Asare, Thomas Lumley, et al. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, 77(2):121–131, 2010.

[6] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[7] Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

[8] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895.

[9] Yu Qian, Yue Liu, John Campbell, Elizabeth Thomson, Y Megan Kong, and Richard H Scheuermann. Fcstrans: an open source software system for fcs file conversion and data transformation. *Cytometry Part A*, 81(5):353–356, 2012.

[10] Mario Roederer. Compensation in flow cytometry. *Current Protocols in Cytometry*, pages 1–14, 2002.

[11] Claude E Shannon. The mathematical theory of communication. 1963. *MD computing: computers in medical practice*, 14(4):306–317, 1996.

[12] Howard M Shapiro. *Practical flow cytometry*. John Wiley & Sons, 2005.

[13] Vladimir Vapnik and Alexey Chervonenkis. A note on one class of perceptrons. *Automation and remote control*, 25(1), 1964.

[14] Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.