UNIVERSITY OF BRISTOL

DEPARTMENT OF
ENGINEERING MATHEMATICS

Deep Learning Methods in Genomic Medicine

Matthew Ramcharan (Engineering Mathematics)

Project thesis submitted in support of the degree of Master of Engineering

March 25, 2019

Supervisors: Dr I C G Campbell, Engineering Mathematics

Abstract

This is my abstract

Contents

1 Introduction

1.1 Motivation

Somatic mutations are any alteration in cell that will not be passed onto future generations [1]. A somatic mutation in a cell of a fully developed organism can have little to no noticeable effect on the organism itself (often leading to benign growths), however mutations that give rise to cancer are a special case. Cancer arises either from inactivation of tumour suppressor genes, or mutation of a special category of genes called proto-oncogenes, many of which regulate cell division. When mutated, proto-oncogenes enter a state of uncontrolled division and become oncogenes, resulting in a cluster of cells called a tumour. These types of cell division lead to malignant tumours, in which the excessive cell proliferation causes the tumour to spread into surrounding tissues and cause damage.

A common, probably simplistic, model view defines two classes of mutations, 'driver' mutations, i.e. mutations that give a cancer cell a particular selective advantage, and functionally irrelevant 'passenger' mutations. Discovering functionally important mutations, including clear 'drivers' is one goal of genome re-sequencing efforts [2]. To understand the functional contribution of molecular alterations to oncogenesis, response to therapy and evolution of resistance to therapy it is important to have tools that predict the functional implications of mutations as early in the discovery process as possible.

1.2 Dataset

As a dataset, genome sequences are stored as Singular nucleotide polymorphisms, which are the difference in a single DNA building block, called a nucleotide. When SNPs occur within a gene (the coding region) or in a regulatory region near a gene (certain parts of the non-coding regions), they may play a more direct role in disease by affecting the gene's function.

The coding region, the portion of the genome which codes for proteins, accounts for only about 2% of the whole sequence, and it is becoming increasingly evident that non-coding portions of the genome play crucial functional roles in human development and disease [3]. This implies

there is merit to attempting the same methods on data from both the coding and non-coding regions of the human genome.

1.3 Project Aims and Objectives

In this project we focus on prediction of the effects of somatic point mutations leading to amino acid substitutions[4] in the coding and non-coding region of the human cancer genome. These predictions will be assigned a label as to if a point mutation is oncogenic (Likely cancerous) or benign. As such, the problem outlined in this paper is that of a binary classification problem. There are many cancer sequence databases currently being compiled, such as the Cancer Genome Atlas, COSMIC, and the National Cancer Institute and an large aspect of this project is selecting appropriate data to correctly train a cancer predictor, then test it holds up to a variety of data sources.

1.4 Plan of Report

2 Support Vector Machines

Support Vectors Machines have become a well established tool within machine learning. They work well in practice and have now been used across a wide range of applications from recognizing hand-written digits, to face identification, text categorisation, bioinformatics, and database marketing.

An SVM is an abstract learning machine which will learn from a training data set and attempt to generalize and make correct predictions on novel data.

Since the problem outlined in this project is a binary classification problem. SVMs are an applicable method.

For the training data we have a set of input vectors, denoted x_i , with each input vector having a number of component features. These input vectors are paired with corresponding labels, which we denote y_i and there are m such pairs (i = 1, ... , m).

For this project. It is considered that an oncogenic (cancer causing) sample would be $y_i = +1$, benign sample would $y_i = -1$ and the matching x_i are input vectors encoding various genomic features derived from each mutation i

For two classes of well separated data, the learning task amounts to finding a directed hyperplane, that is, an oriented hyperplane such that datapoints on one side will be labelled $y_i = +1$ and those on the other side as $y_i = -1$.

2.1 Proof

Let us consider a binary classification task with datapoints $\mathbf{x_i}(i = 1, ..., m)$ having corresponding labels $y_i = \pm 1$ and let the decision function be:

$$f(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b) \tag{1}$$

where $\cdot$ is the scalar or inner product (so $\mathbf{w} \cdot \mathbf{x} \equiv \mathbf{w}^T\mathbf{x}$).

3 Multi-Task Learning

4 Multi-task Multiple-kernel Learning

take datasets show multi task is better than single task multi task is better than single kernel gives advantage in terms of test accuracy.

point out that mtmkl for cancers in context of Cscape.

4.1 Widmer ratsch original paper

minimise

$$\mathcal{R}(w) + C\mathcal{L}(w), \tag{2}$$

where $\mathcal{L}(w)$ is loss-term measuring training error and $\mathcal{R}(w)$ is regulariser penalising the complexity of the model, w.

When considering multi task. there are several models $w_1, ..., w_T$, where T is the number of tasks.

normally make a joint regularisation term to penalise discrepancy between individual models

$$\mathcal{R}(w_1, ..., w_T) + C\mathcal{L}(w_1, ..., w_T), \tag{3}$$

Often formulate task similarity matrix for regulariser. Paper

5 Conclusion

References

[1] AJF Griffiths, JH Miller, and DT Suzuki. An Introduction to Genetic Analysis. 7th edition. chapter 15. New York: W. H. Freeman, 7th edition, 2000. ISBN 0-7167-3520-2. URL https://www.ncbi.nlm.nih.gov/books/NBK21894/.

[2] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118, sep 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr407. URL http://www.ncbi.nlm.nih.gov/pubmed/21727090http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3177186.

[3] Manel Esteller. Non-coding RNAs in human disease. *Nature Reviews Genetics*, 12(12):861–874, dec 2011. ISSN 1471-0056. doi: 10.1038/nrg3074. URL http://www.ncbi.nlm.nih.gov/pubmed/22094949http://www.nature.com/articles/nrg3074.

[4] Hashem A. Shihab, Julian Gough, David N. Cooper, Ian N. M. Day, and Tom R. Gaunt. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, 29(12):1504–1510, jun 2013. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt182. URL http://www.ncbi.nlm.nih.gov/pubmed/23620363http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3673218https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt182.