Machine learning network-constrained regression of epigenetic data

Sivo V. Daskalov
Corpus Christi College

UNIVERSITY OF CAMBRIDGE

A dissertation submitted to the University of Cambridge in partial fulfilment of the requirements for the degree of Master of Philosophy in Advanced Computer Science

University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom

Email: sivodaskalov@gmail.com

May 9, 2017

Declaration

I Sivo V. Daskalov of Corpus Christi College, being a candidate for the M.Phil in Advanced Computer Science, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

Total word count: 00,000

Signed:

Date:

Abstract

Computational biology often involves working with high-dimensional data. Penalized regression methods are often used on such data, as they can effectively perform feature selection. Several approaches for network-constrained regression have been suggested in literature over the recent years. They use prior knowledge in the form of a network to exploit known relationships between predictors. Synthetic datasets have been generated to do parameter tuning for the various implemented methods.

We suggest an approach for cooperative parameter tuning in the context of multiple alternative methods that share common input and goals. The aim is to tune the different regression methods iteratively, in a way that increases agreement between their coefficients. Neighboring values on the tuning parameter grid are considered for each method and iteration, selecting the set of values that achieves largest correlation with the averaged coefficients of all other methods for the previous iteration. Given enough iterations and granularity of the tuning grids, this process converges.

We also implement a simple approach to aggregate the coefficients produced by the various regression methods. Each predictor is considered relevant if it corresponds to a non-zero coefficient in a certain fraction of the underlying methods. Once a consensus has been reached through this form of voting, simple linear regression is used to fit only the relevant predictors to the data.

The common way of parameter tuning by minimization of the prediction mean squared error is implemented alongside our suggested approach. The comparison is discussed and a set of tuning parameters is assembled for use on real data. Gene methylation and expression data has been processed with the implemented algorithms. A mapping is created that shows methylation of which genes affects the expression levels of each gene.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Epigenetics [1] studies the heritable traits that cannot be explained by changes in the DNA sequence. Examples of epigenetic mechanisms include DNA methylation and histone modification. These mechanisms adjust the expression level of genes [2], which allows organisms to dynamically adapt to changes in the environment.

Disruption of gene expression levels is related to the development of various diseases [3]. For example, the epigenetic deactivation of certain tumor suppressor genes commonly leads to the development of cancer [4]. The expression levels of certain genes can therefore be used as additional tools in early diagnostics of cancer, as prognosis factors and as predictors of response to treatment.

Good understanding of the relationship between DNA methylation and gene expression is important for both cancer prevention and epigenetic treatment. We have used the gene methylation and expression level data discussed in [5] to explore this relationship. One of the goals in this project is to produce a map that shows the methylation of which genes affects the expression levels of each gene.

This is the introduction where you should introduce your work. In general the thing to aim for here is to describe a little bit of the context for your work — why did you do it (motivation), what was the hoped-for outcome (aims) — as well as trying to give a brief overview of what you actually did.

It's often useful to bring forward some "highlights" into this chapter (e.g. some particularly compelling results, or a particularly interesting finding).

It's also traditional to give an outline of the rest of the document, although without care this can appear formulaic and tedious. Your call.

Chapter 2

Background

A more extensive coverage of what's required to understand your work. In general you should assume the reader has a good undergraduate degree in computer science, but is not necessarily an expert in the particular area you've been working on. Hence this chapter may need to summarize some "text book" material.

This is not something you'd normally require in an academic paper, and it may not be appropriate for your particular circumstances. Indeed, in some cases it's possible to cover all of the "background" material either in the introduction or at appropriate places in the rest of the dissertation.

Chapter 3

Related Work

This chapter covers relevant (and typically, recent) research which you build upon (or improve upon). There are two complementary goals for this chapter:

1. to show that you know and understand the state of the art; and
2. to put your work in context

Ideally you can tackle both together by providing a critique of related work, and describing what is insufficient (and how you do better!)

The related work chapter should usually come either near the front or near the back of the dissertation. The advantage of the former is that you get to build the argument for why your work is important before presenting your solution(s) in later chapters; the advantage of the latter is that don't have to forward reference to your solution too much. The correct choice will depend on what you're writing up, and your own personal preference.

Chapter 4

Design and Implementation

This chapter may be called something else... but in general the idea is that you have one (or a few) "meat" chapters which describe the work you did in technical detail.

Chapter 5

Evaluation

For any practical projects, you should almost certainly have some kind of evaluation, and it's often useful to separate this out into its own chapter.

Chapter 6

Summary and Conclusions

As you might imagine: summarizes the dissertation, and draws any conclusions. Depending on the length of your work, and how well you write, you may not need a summary here.

You will generally want to draw some conclusions, and point to potential future work.

Bibliography

[1] Robin Holliday. Epigenetics: a historical overview. *Epigenetics*, 1(2):76–80, 2006.

[2] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254, 2003.

[3] Gerda Egger, Gangning Liang, Ana Aparicio, and Peter A Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–463, 2004.

[4] Manel Esteller. Epigenetics in cancer. *New England Journal of Medicine*, 358(11):1148–1159, 2008.

[5] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61, 2012.