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 15, 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, ordinary least squares estimation 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 map 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 disease 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.

Several methods [6, 7, 8, 9, 10, 11, 12] have been implemented and considered for use with real data. The hyperparameters for each method have been tuned with the use of synthetic datasets as suggested in [8]. This is done because ground truth remains unknown for the relationship between gene methylation and expression.

A novel method of hyperparameter tuning is developed as an alternative to the widely used method of minimizing the cross-validated mean squared test error. In our context we have a bundle of regression methods that operate on the same training data and share a common goal - to correctly identify the relationship between predictor and target variables. Instead of tuning the various regression methods independently, our approach performs cooperative hyperparameter tuning on all methods simultaneously. It uses an iterative algorithm to increase the similarity of estimated coefficients for the various methods by tuning their hyperparameters. For each method and iteration, the method's parameter grid neighborhood is searched for a set of parameters that maximizes the correlation between its estimated coefficients and the averaged estimates of all other methods for the previous iteration. When this process converges a set of parameters is defined for each method that maximizes the overall agreement across the whole set of methods.

The various regression methods discussed in this project minimize different cost functions. As a result, each method exhibits specific strengths and weaknesses. The relative performance of the methods depends on the dataset used. For this reason it is impossible to predict their effectiveness on an arbitrary real data set with no access to ground truth. We have developed a simple way to merge the estimation results of the method bundle. It is designed to balance the behavior of any individual method. Each of the predictor variables is considered important if it has non-zero coefficients in a fraction of the methods above a given threshold. This approach for variable selection in practice implements a voting system where each predictor must achieve a certain electoral threshold to be selected. Ordinary least squares estimation is then performed only using the set of selected variables.

The following chapter discusses the details of each linear regression approach. After introducing the synthetic dataset generation process, we define our parameter tuning method and compare the results with those from the standard approach. Next we present the real dataset used for exploration of the relationship between gene methylation and expression. After introducing of our result merging approach we present and discuss the results.

Chapter 2

Background and Related Work

This chapter briefly reviews the main concepts of linear regression. We describe in detail the various methods for penalized regression found in literature and used in this project.

2.1 Linear Regression

Let us consider an entity with a number of scalar measurable (observable) properties, for example temperature, weight, and size. We can define a matrix X of n rows and p columns, such that each column contains the observed values of a particular property and each row represents an independent observation of values for all properties. Let us also define a vector y of length n containing the corresponding observed values of an arbitrary property of interest.

Linear regression is an approach for modeling the relationships between a scalar dependent (target) variable y and a number of explanatory variables (predictors) $X_1, ..., X_p$. It assumes that this relationship is linear and assigns a regression coefficient β_i to each predictor X_i, as well as a constant (offset) term β_0. The linear regression model takes the form shown in 2.1

$$y_i = \beta_0 1 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i, \quad for\ i = 1, 2, ..., n \tag{2.1}$$

where ϵ_i represents noise, capturing all external factors influencing the target values, such as inaccuracy of measurement. The error ϵ_i introduces cannot be predicted or reduced.

2.2 Ordinary Least Squares Estimation

Ordinary least squares (OLS) is a method of estimating the unknown parameters β in a linear regression model. It aims to minimize the sum of squared deviations of the observed values from the model prediction (2.2), also called residual sum of squares (RSS).

$$L(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \tag{2.2}$$

The parameter estimate $\hat{\beta}$ for the linear regression model is obtained as shown in equation 2.3 through the minimization of the objective function $S(\beta)$.

$$\hat{\beta} = argmin_{\beta \in R}\ S(\beta) = L(\beta) \tag{2.3}$$

2.3 Penalized Regression

Penalized regression methods introduce a penalty $P(\beta)$ to the objective function $S(\beta)$ in addition to the loss function $L(\beta)$. P penalizes values of the unknown parameters that are considered unrealistic in the current context, which is done to obtain a more meaningful estimation. One or more regularization parameters λ_i can be used to balance the effect of any introduced penalties. The general form of penalized regression is shown in equation 2.4.

$$S(\beta) = L(\beta) + \lambda P(\beta) \tag{2.4}$$

2.4 Lasso

least absolute shrinkage and selection operator

$$S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| \tag{2.5}$$

$alpha = \lambda_1$

2.5 Elastic Net

$$S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sqrt{\sum_{i=1}^{p} \beta_i^2} \tag{2.6}$$

$alpha = \lambda_1 + \lambda_2$ and $l1\_ratio = \lambda_1/(\lambda_1 + \lambda_2)$

2.6 Grace

$$S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{u\sim v} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v) \tag{2.7}$$

2.7 aGrace

$$S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{u\sim v} \left(\frac{sign(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{sign(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}}\right)^2 w(u,v) \tag{2.8}$$

2.8 GBLasso

This method was initially suggested in [10]. Pan, Xie and Shen considered the objective function shown in equation 2.9.

$$S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda 2^{1/\gamma'} \sum_{u\sim v} \left(\frac{|b_u|^\gamma}{w_u} - \frac{|b_v|^\gamma}{w_v}\right)^{1/\gamma}, \tag{2.9}$$

where w_i is a weight function attributed to each node. Three types of functions, dependent on the node's degree d_i and/or γ, were considered by the authors: $w_i = d_i^{(\gamma+1)/2}$, $w_i = d_i$ and $w_i = d_i^\gamma$.

A simplification of the penalty function is presented in [12]. The authors have selected a node weight function of $w_i = d_i^{\gamma/2}$ and the penalty multiplier $\lambda 2^{1/\gamma'}$ has been transformed to λ. The simplified objective function is shown in 2.10.

$$S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda \sum_{u\sim v} \left[\left(\frac{|b_u|}{\sqrt{d_u}}\right)^\gamma - \left(\frac{|b_v|}{\sqrt{d_v}}\right)^\gamma\right]^{1/\gamma} \tag{2.10}$$

2.9 Linf and aLinf

The authors of [11] continued the study presented in section 2.8. They noted that as $\gamma \to \infty$ the penalty 2.10 becomes 2.11.

$$S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda \sum_{u\sim v} \max\left(\frac{|\beta_u|}{\sqrt{d_u}}, \frac{|\beta_v|}{\sqrt{d_v}}\right) \tag{2.11}$$

Luo, Pan and Shen also suggest an equivalent formulation of the penalized estimation in 2.10 as the constrained minimization problem shown in 2.12.

$$
\begin{aligned}
S(\beta) = &\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \\
&\textit{subject to} \sum_{u\sim v}\left[\left(\frac{|b_u|}{\sqrt{d_u}}\right)^{\gamma} - \left(\frac{|b_v|}{\sqrt{d_v}}\right)^{\gamma}\right]^{1/\gamma} \leq C
\end{aligned} \tag{2.12}
$$

As previously shown, under $\gamma \to \infty$ equation 2.12 transforms to 2.13.

$$
\begin{aligned}
S(\beta) = &\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \\
&\textit{subject to} \sum_{u\sim v}\max\left(\frac{|\beta_u|}{\sqrt{d_u}}, \frac{|\beta_v|}{\sqrt{d_v}}\right) \leq C
\end{aligned} \tag{2.13}
$$

The authors suggest an additional modification to reduce bias in the parameter estimates of the standard Linf method. They propose a two-step approach ...

$$
P(\beta) = \lambda\sum_{u\sim v}\left|\frac{sign(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{sign(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}}\right| \tag{2.14}
$$

Which produces the following constrained minimization problem:

$$
\begin{aligned}
S(\beta) = &\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \\
&\textit{subject to} \sum_{u\sim v}\left|\frac{sign(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{sign(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}}\right| \leq E
\end{aligned} \tag{2.15}
$$

2.10 TTLP and LTLP

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.

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 3

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 4

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 5

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.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[7] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[8] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[9] Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4(3):1498, 2010.

[10] Wei Pan, Benhuai Xie, and Xiaotong Shen. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66(2):474–484, 2010.

[11] Chong Luo, Wei Pan, and Xiaotong Shen. A two-step penalized regression method with networked predictors. *Statistics in biosciences*, 4(1):27–46, 2012.

[12] Sunkyung Kim, Wei Pan, and Xiaotong Shen. Network-based penalized regression with application to genomic data. *Biometrics*, 69(3):582–593, 2013.