Computational Statistics and Data Analysis 97 (2016) 60–70

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Natural coordinate descent algorithm for L1-penalised regression in generalised linear models

Tom Michoel

The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK

ARTICLE INFO

Article history:
Received 15 July 2014
Received in revised form 25 August 2015
Accepted 18 November 2015
Available online 28 November 2015

Keywords:
Penalised regression
Generalised linear model
Coordinate descent algorithm
Logistic regression

ABSTRACT

The problem of finding the maximum likelihood estimates for the regression coefficients in generalised linear models with an ℓ_1 sparsity penalty is shown to be equivalent to minimising the unpenalised maximum log-likelihood function over a box with boundary defined by the ℓ_1-penalty parameter. In one-parameter models or when a single coefficient is estimated at a time, this result implies a generic soft-thresholding mechanism which leads to a novel coordinate descent algorithm for generalised linear models that is entirely described in terms of the natural formulation of the model and is guaranteed to converge to the true optimum. A prototype implementation for logistic regression tested on two large-scale cancer gene expression datasets shows that this algorithm is efficient, particularly so when a solution is computed at set values of the ℓ_1-penalty parameter as opposed to along a regularisation path. Source code and test data are available from http://tmichoel.github.io/glmnat/.

1. Introduction

In high-dimensional regression problems where the number of potential model parameters greatly exceeds the number of training samples, the use of an ℓ_1 penalty which augments standard objective functions with a term that sums the absolute effect sizes of all parameters in the model has emerged as a hugely successful and intensively studied variable selection technique, particularly for the ordinary least squares (OLS) problem (e.g. Tibshirani, 1996, Osborne et al., 2000a, Osborne et al., 2000b, Efron et al., 2004, Zou and Hastie, 2005, Johnstone and Titterington, 2009, Friedman et al., 2010, El Ghaoui et al., 2012, Tibshirani et al., 2012 and Tibshirani, 2013). Generalised linear models (GLMs) relax the implicit OLS assumption that the response variable is normally distributed and can be applied to, for instance, binomially distributed binary outcome data or Poisson distributed count data (Nelder and Wedderburn, 1972). However, the most popular and efficient algorithm for ℓ_1-penalised regression in GLMs uses a quadratic approximation to the log-likelihood function to map the problem back to an OLS problem and although it works well in practice, it is not guaranteed to converge to the optimal solution (Friedman et al., 2010). Here it is shown that calculating the maximum likelihood coefficient estimates for ℓ_1-penalised regression in generalised linear models can be done via a coordinate descent algorithm consisting of successive soft-thresholding operations on the *unpenalised* maximum log-likelihood function without requiring an intermediate OLS approximation. Because this algorithm can be expressed entirely in terms of the natural formulation of the GLM, it is proposed to call it the *natural coordinate descent algorithm*.

To make these statements precise, let us start by introducing a response variable $Y \in \mathbb{R}$ and predictor vector $X \in \mathbb{R}^p$. It is assumed that Y has a probability distribution from the exponential family, written in canonical form as

$$p(y \mid \eta, \phi) = h(y, \phi) \exp\Big(\alpha(\phi)\big\{y\eta - A(\eta)\big\}\Big),$$

E-mail address: tom.michoel@roslin.ed.ac.uk.

http://dx.doi.org/10.1016/j.csda.2015.11.009
0167-9473/

where $\eta \in \mathbb{R}$ is the natural parameter of the distribution, ϕ is a dispersion parameter and $h, \alpha > 0$ and A convex are known functions. The expectation value of Y is a function of the natural parameter, $E(Y) = A'(\eta)$, and linked to the predictor variables by the assumption of a linear relation $\eta = X^T\beta$, where $\beta \in \mathbb{R}^p$ is the vector of regression coefficients. It is tacitly assumed that $X_1 \equiv 1$ such that β_1 represents the intercept parameter. Suppose now that we have n observation pairs (x_i, y_i) (with $x_{i1} = 1$ fixed for all i). The minus log-likelihood of the observations for a given set of regression coefficients β under the GLM is given by

$$H(\beta) = \frac{1}{n}\sum_{i=1}^{n} A(x_i^T\beta) - y_i(x_i^T\beta) \equiv U(\beta) - w^T\beta, \tag{1}$$

where any terms not involving β have been omitted, $U(\beta) = \frac{1}{n}\sum_{i=1}^{n} A(x_i^T\beta)$ is a convex function, $w = \frac{1}{n}\sum_{i=1}^{n} y_i x_i \in \mathbb{R}^p$, and the dependence of U and w on the data (x_i, y_i) has been suppressed for notational simplicity. In the penalised regression setting, this cost function is augmented with ℓ_1 and ℓ_2 penalty terms to achieve regularity and sparsity of the minimum-energy solution, i.e. H is replaced by

$$H(\beta) = U(\beta) - w^T\beta + \lambda\|\beta\|_2^2 + \mu\|\beta\|_1, \tag{2}$$

where $\|\beta\|_2 = (\sum_{j=1}^{p} |\beta_j|^2)^{\frac{1}{2}}$ and $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ are the ℓ_2 and ℓ_1 norm, respectively, and λ and μ are positive constants. The ℓ_2 term merely adds a quadratic function to U which serves to make its Hessian matrix non-singular and it will not need to be treated explicitly in our analysis. Furthermore a slight generalisation is made where instead of a fixed parameter μ, a vector of predictor-specific penalty parameters μ_j is used. This allows for instance to account for the usual situation where the intercept coefficient is unpenalised ($\mu_1 = 0$). The problem we are interested in is thus to find

$$\hat{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} H(\beta), \tag{3}$$

with H a function of the form

$$H(\beta) = U(\beta) - w^T\beta + \sum_{j=1}^{p} \mu_j|\beta_j|, \tag{4}$$

where $U : \mathbb{R}^p \to \mathbb{R}$ is a smooth convex function, $w \in \mathbb{R}^p$ is an arbitrary vector and $\mu \in \mathbb{R}^p$, $\mu \succcurlyeq 0$ is a vector of non-negative parameters. The notation $u \succcurlyeq v$ is used to indicate that $u_j \geq v_j$ for all j and likewise the notation $u \cdot v$ will be used to indicate elementwise multiplication, i.e. $(u \cdot v)_j = u_j v_j$. The maximum of the *unpenalised* log-likelihood, considered as a function of w, is of course the Legendre transform of the convex function U,

$$L(w) = \max_{\beta\in\mathbb{R}^p}\Big\{w^T\beta - U(\beta)\Big\},$$

and the unpenalised regression coefficients satisfy

$$\hat{\beta}_0(w) = \operatorname*{argmax}_{\beta\in\mathbb{R}^p}\Big\{w^T\beta - U(\beta)\Big\} = \nabla L(w),$$

where ∇ is the usual gradient operator (see Lemma 1 in Appendix A.1). This leads to the following key result:

Theorem 1. *The solution $\hat{\beta}(w, \mu)$ of*

$$\hat{\beta}(w, \mu) = \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\Big\{U(\beta) - w^T\beta + \sum_{j=1}^{p} \mu_j|\beta_j|\Big\} \tag{5}$$

is given by

$$\hat{\beta}(w, \mu) = \hat{\beta}_0\big(\hat{u}(w, \mu)\big) = \nabla L\big(\hat{u}(w, \mu)\big),$$

where $\hat{u}(w, \mu)$ is the solution of the constrained convex optimisation problem

$$\hat{u}(w, \mu) = \operatorname*{argmin}_{\{u\in\mathbb{R}^p\,:\,|u-w|\preccurlyeq\mu\}} L(u). \tag{6}$$

Furthermore the sparsity patterns of $\hat{\beta}$ and $\hat{u} - w + \operatorname{sgn}(\hat{\beta}) \cdot \mu$ are complementary,

$$\hat{\beta}_j(w, \mu) \neq 0 \Leftrightarrow \hat{u}_j(w, \mu) = w_j - \operatorname{sgn}(\hat{\beta}_j)\mu_j.$$