Fitting Generalized Additive Models for very large datasets with Apache Spark

Chapter Excerpt

Kai Thomas Brusch

24.12.2015

Summary This document contains the main introductory chapters for my bachelor thesis titled "Fitting General Additive Models for very large datasets with Apache Spark".

Contact kai.brusch@gmail.com

Location Hochschule für Angewandte Wissenschaften Hamburg

Department Dept. Informatik

Examiner Prof. Dr. Michael Köhler-Bußmeier

Second examiner Dipl.-Math. Markus Schmaus

Contents

1 Linear models

1.1 Introduction to linear models

Linear models are statistical models in which an univariate response is modeled as the sum of a 'linear predictor' and a zero mean random error term. The linear predictor depends on some predictor variables y, measured with the response variable x, and some unknown parameters β plus an error term ϵ, which must be estimated. This process is formally stated for a given row i of data as:

$$y_{\mathrm{i}} = \beta x_{\mathrm{i}} + \varepsilon_{\mathrm{i}} \tag{1}$$

The are many choices for β and finding the best possible β stands at the heart of the following chapter. A key feature of linear models is that the linear predictor depends linearly on these parameters. Statistical inference with such models is usually based on the assumption that the response variable has a normal distribution. Linear models are used widely in most branches of science.

1.2 Example of a linear model

I am an example of a simple linear model. I explain some variable y as linear combination of a model matrix and an estimated coefficent.

1.3 Ordinary least square estimation of β

We are now looking at methods of finding β. Ideally we want to choose a β that produces a line through our data points with minimal distance between our points and our estimated line. More precicely we are looking to estimate a β that minimizes the squared distance between an estimated β times the given x and y_i. We are squaring the distance to normalize negative and positive differences. This distance formaly describes as S:

$$\mathbf{S} = \sum_{i=1}^{n} (y_{\mathrm{i}} - x_{\mathrm{i}}\beta)^2 \tag{2}$$

The close our S gets to 0 the better our line fits the data. The Markov-Gauss Theorem states that the minimization of S yields $\widehat{\beta}$ which is the best possible estimation for b. This shall be discussed n the next chapter. 1 represents the univariate case where y is explained with only one variable. The process of minimizing S is commonly refered to as ordinary least squares (OLS). This model can be extended to multiple independent variables yielding in a scalar matrix form.

$$y = \mathbf{X}\beta + \epsilon \tag{3}$$

The dependent variable vector y is the linear combination of model matrix X and the vector of unknown parameter β. Finding the best possible estimator $\widehat{\beta}$ for 3 is formalized stated as the minimal length of the distance between y and Xβ:

$$\|y - X\beta\|^2 \tag{4}$$

There are generally two ways to think about estimating β. From the Calculus point of view we can see this as a minimization problem of a function with two parameters: S and β. Hence we minimize S in respect to β by taking the partial derivative.

$$\frac{\partial S}{\partial \beta} = -\sum_{i=1}^{n} 2x_i(y_i - x_i\beta) \tag{5}$$

Rewriting 5 will yield:

$$-\sum_{i=1}^{n} 2x_i(y_i - x_i\widehat{\beta}) = 0 \tag{6}$$

$$-\sum_{i=1}^{n} x_i y_i - \widehat{\beta}\sum_{i=1}^{n} x_i^2 = 0 \tag{7}$$

$$\widehat{\beta} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 \tag{8}$$

Minimizing S w.r.t. β is a theoretically reasonable approach to estimating β

From the Linear Algebra point of view we can see our data as an underdetermined system of equations.

1.4 Gauss-Markov Theorem

What is so special about least squares? It is the best, unbiased estimation of β. This shall be shown!

1.5 Estimation of $\widehat{\beta}$ by orthogonal decomposition

Projection to orthogonal space to find *beta*.

1.6 Geomoery of linear models

Steal plots from GAM book

2 Generalized Linear Models

2.1 Introduction to Generalized Linear Models

Generalized linear models (GLMs) relax the strict linearity assumption of linear models, by allowing the expected value of the response to depend on a smooth monotonic function of the linear predictor. Similarly the assumption that the response is normally distributed is relaxed by allowing it to follow any distribution from the exponential family (normal, Poisson, Binomial, Gamma etc.). While OLS was sufficient for estimating β for normal distributed data we have to generalize this notion to acount for an arbitrary amount of distribution parameters.

2.2 Example of a Generalized Linear Models

I am a GLM

2.3 Maximum likelihood estimation

OLS is not sufficent to account for the expotential family. MLE generalizes OLS to account for an fixed number of distribution parameters. We estimate those then.

2.4 Fitting generalized linear models

Penelized Iterative Reweighted Least Square estimation

2.5 Geomoery of Generalized Linear Models

Steal plots from GAM book

3 Generalized Additive Models

3.1 Introduction to Generalized Additive Models

Generalized Additive Models (GAMs) extends the GLM by specifying the linear prediction in terms of the summation of smooth functions. This allows for a more flexible modeling of the influence for each explanatory variable. The gained flexibility comes at the cost of additional questions concerning the smooth function:

GAMs are fomally described by the following equation:

$$g(\mu_\mathrm{i}) = \mathbf{X}_\mathrm{i}\Theta + f_1(x_{1\mathrm{i}}) + f_2(x_{2\mathrm{i}}) + f_3(x_{3\mathrm{i}}, x_{4\mathrm{i}})... \tag{9}$$

9 explains y_i as the model matrix for this row and the smooth functions $f_j(x_1 j)$ of the x values for this row. X_i is a row of the model matrix with parametric component θ. Unlike the linear model we can now specify a smooth function for each explanatory variable. This proves to be way more flexible than only allowing for a constant influence per explanatory variable. The natural question that arises now are: How do I find proper smoothing functions? Finding the right smooth function stands at the heart of GAM fitting and can be best illustrated in the univariate case 10.

3.2 Generalized Additve Model example

WOODS HELP ME!!!

3.3 Smoothing Functions

$$y_i = f(x_i) + \epsilon_i \tag{10}$$

Smooth functions form a vector space, which can be approximated using a linear basis.Only allowing linear basis allow us to heavily leverage the theory already developed for linear models and S as the optimal model fit. For the sake of illustration we assume that 10 can be rewriten as the following equation if $b_i(x)$ is the ith basis function:

$$f(x) = \sum_{i=1}^{q} b_\mathrm{i}(x)\beta_\mathrm{i} \tag{11}$$

In 11 we already know $f()$ is linear in regard to 11. We now have to specify a basis function to represent bi. We can choose from many basis functions for bi, each with advantages and disadvantages. A common choice however is a fourth order polynomial basis function. 11 represented by a fourth order polynomial yields the following model:

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5 \tag{12}$$

Applying 12 to 10 we get the modeling of y_i as the sum of smoothing functions.

$$y_i = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + x_i^3\beta_4 + x_i^4\beta_5 + \epsilon_i \tag{13}$$

3.4 Regression Splines

3.5 Smoothing Parameter estimation

3.6 Fitting Generalized Additive Models

4 Matrix Algebra

4.1 Orthogonal Matrices

We are concerned with matrices that do have orthogonal column vectors

4.2 Cholesky decomposition

We want to decompose a matrix A in two parts

4.3 QR decomposition

We want to decompose a matrix A in two parts, one orthonormal and one upper triangular, X = QR