Fitting Generalized Additive Models for very large datasets with Apache Spark

Chapter Excerpt

Kai Thomas Brusch

24.12.2015

Summary This document contains three 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. 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 β. 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).

There are two ways to think about estimating β, thinking off OLS in terms of calulus on functions allows us to compute $\widehat{\beta}$ for the univariate case while computing $\widehat{\beta}$ for multiple independant variables requires linear algebra. From the calculus perspective we can see OLS as function with two parameters: S and β. Minimizing S equates to taking the partial derivative of S with repsect to β. This is the most common approach and offers insight by stating S as the following equation:

$$\frac{\partial S}{\partial \beta} = -\sum_{i=1}^{n} 2x_i(y_i - x_i\beta) \tag{3}$$

Rewriting the partial dereviative to yiel $\widehat{\beta}$ gives a very good idea of β

$$-\sum_{i=1}^{n} 2x_i(y_i - x_i\widehat{\beta}) = 0 \tag{4}$$

$$-\sum_{i=1}^{n} x_i y_i - \widehat{\beta} \sum_{i=1}^{n} x_i^2 = 0 \tag{5}$$

$$\widehat{\beta} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 \tag{6}$$

Minimizing S w.r.t. β to compute $\widehat{\beta}$ is a reasonable approach when dealing with one independant variable. However, almost all relevant applications involve much more than one independant variable and require linear algebra. To estimate $\widehat{\beta}$ with OLS for multiple independant variables involves rephrasing the questions in terms of linear algebra. First we have to state the process of finding β as a linear combination problem whichs equates to asking what linear combination of column vectors of our model matrix X and our vector β of unknown coefficents yields the vector of y. Formally:

$$X\beta = y \tag{7}$$

Now that we have restated OLS as a matrix problem we can apply some linear algebra to find $\widehat{\beta}$. First we can restate the problem of finding β as finding the unkown vector β equates to finding a linear combination of our column vectors of X with β that result in y. Finding this linear combination is highly dependant on the properties of X. Taking a look at relevant model matricies X will show that they almost exclusively consist of m rows and n columns with $m > n$. Being $m > n$ implies that the matrix is not symetric and not invertible. Given that X has more rows than columns we can think of $X\beta = y$ as a system of equations with more equations than variables which causes this sytem of equation to have no solution. We have to stress the fact that the sytem of equations of 7 does not have a solution because there is no possible selecton for β that lies in the column vector space of X. Whilst the nature of X makes finding β impossible we can find an estimation $\widehat{\beta}$ of β by projecting it back into the column space of X. The solution becomes to multiply by A^TThe proper projection p is defined as $p = X\widehat{\beta}$. The process of projecting of finding the projection involves multiplying by the transpose of the model matrix yielding:

$$X^T X\widehat{\beta} = X^T y \tag{8}$$

This projection comes however at the cost of an error term:

$$y = X\widehat{\beta} + \epsilon \tag{9}$$

$$\epsilon = y - X\widehat{\beta} \tag{10}$$

The questions that now arises is: how to we make $\epsilon = y - X\widehat{\beta}$ as small as possible? Using algebra we can split the vector β into two parts. One part in the column space is our projection p and the perpendiclar part in the nullspace of A^T which the ϵ. It is essential to remember that the column space is always perpendicular to the nullspace of A^T. The solution to $X\beta = p$ leaves the least possible error ϵ, returning to the previously stated method of least squares, but this time in the world of linear algebra:

$$\|X\beta - y\|^2 = \|X\beta - p\|^2 + \|\epsilon\|^2 \tag{11}$$

This is the law $c^2 = a^2 + b^2$ for a right angle. The vector $X\beta - p$ in the column space is perpendicular to ϵ in the nullspace of A^T. We reduce $X\beta - p$ to zero choosing β to be $\widehat{\beta}$, this leaves us with the smallest possible error vector ϵ.

The projection leaves us with an invertible matrix that can be solved by usual elimination.

The least squares solution $\widehat{\beta}$ makes $\epsilon = X\beta$ as small as possible.

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 Estimating $\widehat{\beta}$ with orthogonal decomposition

The previously suggested method lends a great tool to think about the univariate least squares technique. But the suggested methods is rarely applied Any m by n matrix X with independant columns can be factored into QR. The m by n matrix Q has orthonormal columns and the square matrix R is upper triangular with positive diagonal. X^TX equals $R^TQ^TQR = R^TR$ simplifying the leas squares equation to $Rx = Q^T\beta$; this additional simplicity allows us to restate the problem of least squares for matrixes using QR decomposition as:

$$R^TR\widehat{\beta} = R^TQ^Ty \ or \ R\widehat{\beta} = Q^Ty \ or \ \widehat{\beta} = R^{-1}Q^Ty \tag{12}$$

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:

$$g(\mu_{\mathrm{i}}) = \mathbf{X}_{\mathrm{i}}\Theta + f_1(x_{1\mathrm{i}}) + f_2(x_{2\mathrm{i}}) + f_3(x_{3\mathrm{i}}, x_{4\mathrm{i}})... \tag{13}$$

?? explains y_i as the model matrix for this row and the smooth functions $f_j(x_1j)$ 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 **??**.

3.2 Generalized Additve Model example

WOODS HELP ME!!!

3.3 Smoothing Functions

$$y_i = f(x_i) + \epsilon_i \tag{14}$$

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 **??** can be rewriten as the following equation if $b_i(x)$ is the ith basis function:

$$f(x) = \sum_{i=1}^{q} b_{\mathrm{i}}(x)\beta_{\mathrm{i}} \tag{15}$$

In **??** we already know $f()$ is linear in regard to **??**. 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. **??** represented by a fourth order polynomial yields the following model:

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5 \tag{16}$$

Applying **??** to **??** we get the modeling of y_i as the sum of smoothing functions.

$$y_i = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + x_i^3\beta_4 + x_i^4\beta_5 + \epsilon_i \tag{17}$$

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, severly concerned

4.2 QR decomposition

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

5 References

[Wood, 2006] Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

[Wood et al., 2015] Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

[Zaharia et al., 2010] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA. USENIX Association.