gamdist: Generalized Additive Models in Python

Bob Wilson

August 29, 2018

Abstract

TBD

Contents

1 Introduction

This paper introduces a Python library called `gamdist`, which uses a distributed optimization technique called the Alternating Direction Method of Multipliers (ADMM) to fit a special type of regression model called a Generalized Additve Model (GAM) to data.

Outline of Paper In §2 we describe Generalized Additive Models. In §3 we describe the Alternating Direction Method of Multipliers and how it may be used to fit GAMs. In §4, we describe the architecture of the library, including relevant implementation details.

2 Generalized Additive Models

The primary goal of `gamdist` is the estimation of certain aspects of the joint distribution of a collection of one or more random variables X called *features* and a random variable Y we will call the *response*. Specifically, we are interested in the conditional distribution of $Y \mid X$. Perhaps the simplest approach is the linear model, which assumes $Y \mid X = x \sim \mathcal{N}(\mu(x), \sigma^2)$, where $\mu(x) = \nu(x)^T\beta$. This notation captures three key assumptions of the linear model. First, the conditional distribution is Gaussian for all values of X. Second, the mean of the distribution depends on the features X in a fairly specific way discussed below. Third, the variance is the same for all values of X. These assumptions are all loosened in various generalizations of the linear model used in `gamdist`.

If we choose $\nu(x) = x$, then the assumption is that $\mu(x) = x^T\beta$, and the mean depends linearly on the features; however, the *linear* in *linear model* refers to the dependence on β, not on the features. It is common to include a constant term in $\nu(x)$ to account for an affine dependency between the features and the response. For example, $\nu(x) = \begin{bmatrix}1 & x_1 & x_2 & \cdots\end{bmatrix}^T$. We might incorporate quadratic terms to capture nonlinear dependencies including interactions, such as $\nu(x) = \begin{bmatrix}x_1 & x_2 & x_1^2 & x_1 \cdot x_2 & x_2^2 & \cdots\end{bmatrix}^T$. Or we might include more exotic transformations of the features, like $\nu(x) = \begin{bmatrix}\log(x_1) & \sin(x_2) & \cdots\end{bmatrix}$. These models are nonlinear in the features, but linear in the parameters β; however, the linear model will only incorporate transformations that we explicitly include.

Another common situation is when some or all of the features are categorical. For example, consider a model with a single feature corresponding to a person's favorite color, and suppose choices are limited to red, green, and blue. The model would consist of the average responses for people who prefer any particular color. We might support such a model by defining

$$\nu(x) = \begin{cases} \begin{bmatrix}1 & 0 & 0\end{bmatrix}^T & \text{if } x = \text{red} \\ \begin{bmatrix}0 & 1 & 0\end{bmatrix}^T & \text{if } x = \text{green} \\ \begin{bmatrix}0 & 0 & 1\end{bmatrix}^T & \text{if } x = \text{blue.} \end{cases}$$

Alternative approaches to encoding categorical variables are common and useful in different circumstances. We see that the simple linear model is applicable to a wide range of problems,

even those that may not appear linear at first glance.

Fitting a linear model to a set of observations is called linear regression, and is accomplished by solving a least squares optimization problem. This is an example of maximum likelihood estimation (MLE), itself a special case of maximum a priori (MAP) estimation, which is the unifying approach used throughout `gamdist`. It is worth formulating this optimization problem so that we may see how it evolves as we consider more general scenarios.

Suppose we have n observations of the form $\{(x^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$. We assume these observations are independent and drawn from the same (unknown) joint distribution.[1] Under the assumptions of the linear model, $Y \mid X = x^{(i)} \sim \mathcal{N}(\mu(x^{(i)}), \sigma^2)$. The likelihood of a particular observation is

$$\mathcal{L}(\beta; x^{(i)}, y^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \nu\left(x^{(i)}\right)^T \beta\right)^2\right).$$

Note that by convention the likelihood is interpreted as a function of β parameterized by the observation $(x^{(i)}, y^{(i)})$. The likelihood of the entire set of observations is the product of the likelihoods of the individual observations: $\mathcal{L}(\beta; x, y) = \Pi_{i=1}^{n} \mathcal{L}(\beta; x^{(i)}, y^{(i)})$. The log-likelihood is the sum of the log-likelihoods of the individual observations:

$$\ell(\beta; x, y) = \log \mathcal{L}(\beta; x, y) = -n/2 \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \nu\left(x^{(i)}\right)^T \beta\right)^2.$$

Maximizing the likelihood is the same as maximizing the log-likelihood, and if we are only interested in estimating β, this is equivalent to the problem

$$\text{minimize} \quad \sum_{i=1}^{n} \left(y^{(i)} - \nu\left(x^{(i)}\right)^T \beta\right)^2,$$

where the variable is β and $y^{(i)}$ and $\nu(x^{(i)})$ are data. An elementary result in optimization theory is that a unique solution exists if and only if $V^T V$ is full rank, where the ith row of V is equal to $\nu(x^{(i)})^T$. In that case, the optimal β satisfies the so-called normal equations:

$$V^T V \hat{\beta} = V^T y.$$

If we assume the model is correct (that is, that the conditional distribution really has the assumed form), exact formulae exist for confidence intervals on the parameters β. If the variance is unknown, it too can be estimated from the data. We can employ hypothesis tests against the null hypothesis that some or all of the components of β are zero. We can apply the resulting model to new data assumed to be drawn from the same joint distribution to compute confidence intervals on the response, $Y \mid X = x_{\text{new}}$ or the mean of this distribution, $\mu(x_{\text{new}})$. These are immensely valuable tools in the analysis of data and the application of

[1] This assumption is loosened in random effects or mixed effects models which are not considered here; see [Str12].

data to predictions. A good reference on linear models is [Wei05]. Useful results are collected in Appendix A.

Even if the assumptions underlying the linear model are correct, if the noise is high, or if the number of features is large relative to the number of observations, we may use regularization to improve both the estimates of β and predictions based on the estimated model. Regularization reduces the sensitivity of the estimates to noise at the expense of introducing bias, and may be thought of as imposing a Bayesian prior on the parameters β. Some of the most common forms of regularization include ridge regression and the lasso [Tib96]. For example, the lasso may be formulated as the problem:

$$\text{minimize} \quad \sum_{i=1}^{n} \left(y^{(i)} - \nu \left(x^{(i)} \right)^T \beta \right)^2 + \lambda \cdot \|\beta\|_1,$$

but this problem does not have a closed-form solution. Moreover, introducing regularization means the formulation is no longer a maximum likelihood estimation problem. Instead, it is a maximum a priori estimation problem. MLE problems have some statistical properties that MAP estimation problems do not possess.

If the constant variance assumption does not hold, but the dependence is known, then Weighted Least Squares.

If the conditional distribution is not Gaussian, other techniques may prove more useful. For example, if the conditional distribution is Laplacian, we may use least absolute deviation regression instead of least squares [BD93]. Like with the lasso, there are no exact formulae for statistical inference in this context and we must settle for an asymptotic or non-parametric approach such as the bootstrap [ET93].

Yet another approach to extending linear models was introduced by [NW72] and discussed in detail in [MN83]. Their formulation extends the linear model in a few ways. The conditional distribution is not assumed to be Gaussian. Common alternatives include the binomial and Poisson distributions. The mean of the distribution is permitted to depend on the features in a more complicated way, via the introduction of a *link function*, g: $\mu(x) = g^{-1}(\eta(x))$, where $\eta(x) = \nu(x)^T \beta$. When $g(x) = x$, this recovers the same relationship between the features and μ assumed in the linear model, but other link functions may be used like the logistic function $g(x) = \log(x/(1-x))$. Finally, the variance is sometimes permitted to depend on x instead of being constant. Such models are called Generalized Linear Models (GLMs). Fitting such models is accomplished via maximum likelihood estimation:

$$\text{minimize} \quad \sum_{i=1}^{n} \ell \left(\beta; x^{(i)}, y^{(i)} \right) + r(\beta),$$

where $r(\beta)$ is a regularization term on β, such as in the lasso. For many choices of distribution family and link function, the corresponding optimization problem is convex.

Introduced by [HT86], Generalized Additive Models (GAMs) extend GLMs by permitting $\eta(x)$ to be a nonparametric function of the features: $\eta(x) = \sum_{i=1}^{p} h_i(x_i)$, where h_i are smooth functions. When $h_i(x_i) = \beta_i x_i$, the linear model is recovered (all of the parametric

dependencies discussed with regards to ν are still possible here of course, but the idea is that the data itself should tell us the form of the relationship). Typically the h_i functions are chosen to be some sort of spline, such as a natural cubic spline. GAMs are also fit via MLE, and many practical problems can be formulated as convex optimization problems.

(Reference GAM, GAMr, Casella and Berger, Seber) regularized GAMs

3 The Alternating Direction Method of Multipliers

4 Software Architecture

Acknowledgments

This is an example of an unnumbered section.

A Properties of the Linear Model

In this section, we state various useful facts (without proof) regarding the linear model. For details, see [Wei05], CB, GAMr.

Suppose $Y \mid X = x \sim \mathcal{N}(\mu(x), \sigma^2)$, where $\mu(x) = \nu(x)^T\beta$ and $\nu(x) \in \mathbf{R}^p$. Let $\{(x^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$ be IID samples drawn from the joint distribution of X and Y. Let V be the matrix whose ith row is $\nu(x^{(i)})$, and assume V^TV is full rank. Let $\hat{\beta} = (V^TV)^{-1}V^Ty$. Then

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(V^TV)^{-1}). \tag{1}$$

Specifically, $\hat{\beta}$ has a multivariate normal distribution, and $\hat{\beta}$ is an unbiased estimate of β. It is also a *consistent* estimate of β, meaning that $\hat{\beta}$ converges in probability to β, as n increases without limit. It is also the best linear unbiased estimate of β: any other linear, unbiased estimates have higher variance than $\hat{\beta}$.

If σ^2 is unknown, it may be estimated from the data. Let

$$\hat{\sigma}^2 = \frac{\|y - V\hat{\beta}\|_2^2}{n-p}. \tag{2}$$

Then $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$. Suppose $\mathbf{Prob}\{\chi^2_{n-p} \in (\ell, u)\} = \alpha$; that is, (ℓ, u) is a confidence interval (at level α) on a χ^2_{n-p} random variable. Then

$$\mathbf{Prob}\left\{\sigma^2 \in \left(\frac{(n-p)\cdot\hat{\sigma}^2}{u}, \frac{(n-p)\cdot\hat{\sigma}^2}{\ell}\right)\right\} = \alpha. \tag{3}$$

A particularly useful case is when $u \to \infty$, corresponding to $\ell = \Phi^{-1}(1-\alpha)$, where Φ is the cumulative distribution function of a χ^2_{n-p} random variable, in which case $\frac{(n-p)\cdot\hat{\sigma}^2}{\Phi^{-1}(1-\alpha)}$ is an upper confidence limit on σ^2, at level α.

We may compute confidence intervals on linear combinations of the components of β by noting that $c^T\hat{\beta} \sim \mathcal{N}(c^T\beta, \sigma^2 \cdot c^T(V^TV)^{-1}c)$, and thus

$$\frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\sigma^2 \cdot c^T(V^TV)^{-1}c}} \sim \mathcal{N}(0,1).$$

Let $\pm z_\alpha$ be the endpoints of a confidence interval on a standard Gaussian random variable at level α. Then

$$c^T\hat{\beta} \pm z_\alpha \cdot \sqrt{\sigma^2 \cdot c^T(V^TV)^{-1}c} \tag{4}$$

are the endpoints of a confidence interval on $c^T\beta$. This formula is only computable when σ^2 is known.

When σ^2 is unknown,

$$\frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\hat{\sigma}^2 \cdot c^T(V^TV)^{-1}c}} \sim t_{n-p},$$

and we may simply replace z_α in Equation (4) by the corresponding value for the Student's t distribution with $n-p$ degrees of freedom.

What about simultaneous confidence intervals on multiple components of β? What about the F-test on the composite hypothesis that all values of β are zero vs the alternative that at least one is nonzero? What about confidence intervals on predictions? What about confidence intervals on the mean prediction? What does hypothesis testing look like? What are the relevant hypotheses? How are p-values calculated? How do we do power calculations?

References

[BD93] D. Birkes and Y. Dodge. *Alternative Methods of Regression.* John Wiley & Sons, Inc., 1993.

[ET93] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, 1993.

[HT86] T. Hastie and R. Tibshirani. Generalized additive models. *Statist. Sci.*, 1(3):297–310, 1986.

[MN83] P. McCullagh and J. Nelder. *Generalized Linear Models.* Chapman & Hall, 1983.

[NW72] J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384, 1972.

[Str12] W. Stroup. *Generalized Linear Mixed Models.* Chapman & Hall, 2012.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[Wei05] S. Weisberg. *Applied Linear Regression.* John Wiley & Sons, Inc., 2005.