

Generalized Additive Models

Overview

Many nonparametric methods do not perform well when there is a large number of independent variables in the model. The sparseness of data in this setting inflates the variance of the estimates. The problem of rapidly increasing variance for increasing dimensionality is sometimes referred to as the “curse of dimensionality.” Interpretability is another problem with nonparametric regression based on kernel and smoothing spline estimates (Hastie and Tibshirani 1990).

To overcome these difficulties, Stone (1985) proposed additive models. These models estimate an additive approximation to the multivariate regression function. The benefits of an additive approximation are at least twofold. First, since each of the individual additive terms is estimated using a univariate smoother, the curse of dimensionality is avoided, at the cost of not being able to approximate universally. Second, estimates of the individual terms explain how the dependent variable changes with the corresponding independent variables.

Hastie and Tibshirani (1990) proposed generalized additive models. These models assume that the mean of the dependent variable depends on an additive predictor through a nonlinear link function. Generalized additive models permit the response probability distribution to be any member of the exponential family of distributions. Many widely used statistical models belong to this general class, including additive models for Gaussian data, nonparametric logistic models for binary data, and nonparametric log-linear models for Poisson data.

The GAM Procedure

The GAM procedure implements the generalized additive model proposed by Hastie and Tibshirani (1990). PROC GAM:

- fits nonparametric or semiparametric additive models
- supports the use of multidimensional data
- supports multiple SCORE statements
- enables you to specify the model degrees of freedom or smoothing parameter

PROC GAM can fit Gaussian, binomial, Poisson, and Gamma distributions. For each distribution, although theoretically more than one link can exist, PROC GAM always uses the canonical link. This is because the difference between link alternatives will be less pronounced for nonparametric models, in light of the flexibility of nonparametric model forms.

The GAM Method

Suppose that Y is a response random variable and X_1, \dots, X_p is a set of predictor variables. A regression procedure can be viewed as a method for estimating how the value of Y depends on the values of X_1, \dots, X_p . The standard linear regression model assumes the expected value of Y has a linear form.

$$E(Y) = f(X_1, \dots, X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

Given a sample of values for Y and X , estimates of $\beta_0, \beta_1, \dots, \beta_p$ are often obtained by the least squares method.

The additive model generalizes the linear model by modeling the expected value of Y as

$$E(Y) = f(X_1, \dots, X_p) = s_0 + s_1(X_1) + \dots + s_p(X_p)$$

where $s_i(X)$, $i = 1, \dots, p$ are smooth functions. These functions are estimated in a nonparametric fashion.

Generalized additive models extend traditional linear models in another way, namely by allowing for a link between $f(X_1, \dots, X_p)$ and the expected value of Y . This amounts to allowing for an alternative distribution for the underlying random variation besides just the normal distribution. While Gaussian models can be used in many statistical applications, there are types of problems for which they are not appropriate. For example, the normal distribution may not be adequate for modeling discrete responses such as counts, or bounded responses such as proportions.

Generalized additive models consist of a random component, an additive component, and a link function relating these two components. The response Y , the random component, is assumed to have a density in the exponential family

$$f_Y(y; \theta; \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

where θ is called the natural parameter and ϕ is the scale parameter. The normal, binomial, and Poisson distributions are all in this family. The quantity

$$\eta = s_0 + \sum_{i=1}^p s_i(X_i)$$

where $s_1(\cdot), \dots, s_p(\cdot)$ are smooth functions defines the additive component. Finally, the relationship between the

mean μ of the response variable and η is defined by a link function $g(\mu) = \eta$. The most commonly used link function is the canonical link, for which $\eta = \theta$.

A combination of backfitting and local scoring algorithms are used in the actual fitting of the model.

Generalized additive models and generalized linear models can be applied in similar situations, but they serve different analytic purposes. Generalized linear models emphasize estimation and inference for the parameters of the model, while generalized additive models focus on exploring data nonparametrically.

Example: Kyphosis Study

Bell and associates (1989) studied the results of corrective spinal surgery commonly performed on children. The specific outcome of interest is the presence (1) or absence (0) of kyphosis (a severe forward flexion of the spine). The available predictor variables are Age in months at the time of the operation, the first vertebrae level involved in the operation (StartVert), and the number of levels involved (NumVert).

The following statements may be used to fit a GAM model.

```
PROC GAM data=kyphosis;
  model kyphosis = spline(NumVert,df=3)
                spline(Age,df=3)
                spline(StartVert,df=3)
                /dist=logist;
  output out=estimate p uclm lclm;
run;
```

The above statements request PROC GAM to fit a logistic additive model with binary dependent variable kyphosis and ordinary independent variables NumVert, Age, and StartVert. Each term is fitted using a B-spline smoother with 3 degrees of freedom.

```

The GAM Procedure
Dependent Variable: Kyphosis
Smoothing Model Component(s): spline(Age) spline(StartVert) spline(NumVert)

Summary of Input Data Set

Number of Observations           83
Number of Missing Observations    0
Distribution                       Binomial
Link Function                      Logit

Iteration Summary and Fit Statistics

Number of local score iterations   9
Local score convergence criterion  2.6635758E-9
Final Number of Backfitting Iterations  1
Final Backfitting Criterion        5.2326788E-9
The Deviance of the Final Estimate  46.610922317

```

Figure 1. Summary Statistics

The first part of the output from PROC GAM (Figure 1) summarizes the input data set and provides a summary for the backfitting and local scoring algorithms. The critical part of the results in output (Figure 2) is the "Analysis of Deviance" table. For each smoothing effect in the model, this table gives a χ^2 -test comparing the deviance between

the full model and the model without this variable. In this case, the analysis of deviance results indicate that the effects of Age and StartVert are significant, while the effect of NumVert is insignificant.

```

The GAM Procedure
Dependent Variable: Kyphosis
Smoothing Model Component(s): spline(Age) spline(StartVert) spline(NumVert)

Regression Model Analysis
Parameter Estimates

Parameter           Parameter Estimate      Standard Error      t Value      Pr > |t|
-----
Intercept            -2.01533                1.45620             -1.38        0.1706
Linear(Age)          0.01213                 0.00794             1.53        0.1308
Linear(StartVert)   -0.18615                0.07628            -2.44        0.0171
Linear(NumVert)     0.38347                 0.19102             2.01        0.0484

Smoothing Model Analysis
Fit Summary for Smoothing Components

Component           Smoothing Parameter      DF      GCV      Num Unique Obs
-----
Spline(Age)         0.999996                  3.000000  328.512864  66
Spline(StartVert)  0.999551                  3.000000  317.646703  16
Spline(NumVert)    0.921758                  3.000000  20.144058   10

Smoothing Model Analysis
Analysis of Deviance

Source              DF      Sum of Squares      Chi-Square      Pr > ChiSq
-----
Spline(Age)         3.00000  10.494369          16.4358         0.0009
Spline(StartVert)  3.00000  5.494968           8.6060         0.0350
Spline(NumVert)    3.00000  2.184518           3.4213         0.3311

```

Figure 2. Model Fit Statistics

References

Bell, D., Walker, J., O'Connor, G., Orrel, J. and Tibshirani, R. (1989), "Spinal Deformation Following Multi-Level Thoracic and Lumbar Laminectomy in Children." Submitted for publication.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

Stone, C.J. (1985), "Additive Regression and Other Nonparametric Models," *Annals of Statistics*, 13, 689–705.

For More Information

For more information about the new GAM procedure and other analytical software in the SAS System, visit the Statistics and Operations Research Community website at www.sas.com/statistics/. The paper "Fitting Generalized Additive Models with the GAM Procedure" will soon be available from this site.