Lecture 8: Additive Models

■ References

Shalizi: Chapter 8

■ Overview

Nonparametric smoothers (such as k-nearest-neighbors, kernel regression, or smoothing splines) are flexible enough to fit most regression functions with vanishing mse as we get more and more data, but these smoothers converge much slower compared to parametric methods (such as linear regression). This is especially true in high dimensions when we have many covariates.

This means that for the same sample size n, one may benefit from a parametric model. On the other hand, parametric model assumptions (such as the linear assumption in linear regression) may add an additional *model bias* or error due to *model misspecifications*; that is, we converge faster but we may converge faster to the wrong solution.

Additive models represent a compromise between standard linear regression models (which are additive and linear in terms of the original covariates) and fully nonparametric models.

We explain how additive models are fit using the backfitting algorithm. We then review inference for linear models and explain how to do inference with additive models.

■ Questions

1. How do you fit additive models?
2. How do you test hypotheses and find interval estimates?

■ Definitions and Notations

1. Additive model

2. Backfitting algorithm

■ TOPICS

1. Fitting additive models

2. Inference for linear smoothers in general.

■ WHAT'S NEXT?

Lecture 9: Inference for linear smoothers.

Nonparametric Review in One Dimension

[Review: Lecture 7 Overview (Nonparametric Smoothing)]

Bias-Variance Tradeoff. When thinking about parametric versus nonparametric regression, it is important to remember the bias-variance tradeoff. Recall: Conditional on $X = x$, the generalization or prediction error decomposes according to

$$\begin{aligned} R(x) = \mathbb{E}[\text{TestErr}(\hat{r}(x))] &\equiv \mathbb{E}\left[\left(Y - \hat{r}(x)\right)^2 \middle| X = x\right] \\ &= \sigma^2 + \text{Bias}(\hat{r}(x))^2 + \text{Var}(\hat{r}(x)). \end{aligned}$$

For a kernel smoother defined as $\hat{r}_n(x) = \sum_i \ell_i(x) Y_i$, where $\ell_i(x) = \frac{K((x-X_i)/h)}{\sum_j K((x-X_j)/h)}$ and K is a kernel function (see Lecture 7), we showed that the risk (or integrated mse plus irreducible error) has the form

Notes:

$$R(h_n) = \sigma^2 + \underbrace{\mathcal{O}(h_n^4)}_{\text{bias}^2} + \underbrace{\mathcal{O}([nh_n]^{-1})}_{\text{variance}},$$

Differentiating the expression (or balancing the bias and variance terms) give the optimal amount of smoothing, $h_* = O(n^{-1/5})$, which corresponds to a risk that decreases at rate

Notes:

$$\mathcal{O}(n^{-4/5})$$

Interestingly, one can prove that this is a fundamental lower bound (so-called minimax rate) in a fully nonparametric setting if r is assumed to have an integrable second derivative.

Generally speaking: A parametric estimator (e.g., linear regression) will have a lower variance than a nonparametric one (e.g., k-nearest-neighbors, kernel regression, or smoothing splines), because it is more restrictive. Meanwhile, the bias depends on the true underlying model. Nonparametric estimators are generally flexible enough that they will have a low bias for a wide range of underlying regression functions, but a parametric estimator (such as linear regression) will only have a low bias if the parametric assumption is approximately correct (i.e., the true model is approximately linear), and can otherwise suffer from high bias.

Of course the expected test error is composed of bias and variance, so both of these quantities are important for predictive performance. In univariate smoothing, i.e., when $X \in \mathbb{R}$, it can often be the case that our considerations for the bias dominate those for the variance (if sample sizes are reasonably large), with the result that we we favor nonparametric methods.

Multiple Nonparametric Regression

Suppose now that the covariate is p-dimensional,

$$\mathbf{X} = (X_1, \dots, X_p)^T.$$

The regression equation takes the form

$$Y = r(X_1, \dots, X_p) + \epsilon. \tag{1}$$

In principle, kernel smoothers and k-nearest neighbors regression carry over to this case easily. Unfortunately, the risk of a nonparametric regression estimator increases rapidly with the dimension p. This is called **the curse of dimensionality**.

For example, recall from Lecture 7 that an (isotropic) multivariate kernel smoother with bandwidth h, has a risk of the form

Notes:

$$R(h) = \sigma^2 + \underbrace{\mathcal{O}(h^4)}_{\text{bias}^2} + \underbrace{\mathcal{O}\left(\frac{1}{nh^p}\right)}_{\text{variance}}$$

which with the optimal choice of bandwidth yields the *nonparametric rate*

Notes:

$$\mathcal{O}\left(n^{-4/(4+p)}\right).$$

Parametric rates are usually just $\mathcal{O}(p/n)$.

again assuming that r has an integrable second derivative.

To make the risk equal to a small number δ we have

$$\delta = \frac{1}{n^{4/(4+p)}}$$

which implies that

$$n = \left(\frac{1}{\delta}\right)^{(p+4)/4}.$$

Thus:

> **To maintain a given degree of accuracy of an estimator, the sample size must increase exponentially with the dimension p.**

So you might need $n = 30000$ points when $p = 5$ to get the same accuracy as $n = 300$ when $p = 1$.

On the other hand, parametric methods like a linear model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \tag{2}$$

typically have a variance that grows merely linearly with p; but the bias can be large. This begs the question: is there some middle ground?

Additive Models

The Additive Compromise

Reading: Shalizi section 8.1

Sometimes, it pays off to use an additive model which falls somewhere in between the parametric and fully nonparametric models. Starting with the linear model in (2), we could simply replace each linear term $X_j\beta_j$ with a smooth nonlinear function $r_j(X_j)$, yielding the **additive model**

Notes:

$$Y = \beta_0 + r_1(X_1) + \cdots + r_p(X_p) + \epsilon.$$

Note that each r_j only depends on the predictor X_j, and that here X_j means the jth column of X, not the jth row.

This is in a sense simpler than the fully nonparametric model (1), because of the restriction that r decomposes into a sum of univariate regression functions over the variables.

Without any restrictions on the functions $r_1, \ldots, r_p$, the additive model is not identifiable, so we usually assume without a loss of generality that

Notes:

$$\begin{aligned} \mathbb{E}[Y] &= \beta_0 \\ \mathbb{E}[r_j(X_j)] &= 0 \qquad \text{for } j = 1, \dots, p \end{aligned}$$

Additive estimates tend to balance the strengths of the fully nonparametric and parametric estimates. That is, additive estimates tend have a lower variance than fully nonparametric ones, and can have a lower bias than parametric ones:

Notes: Even in p dimensions, we have

$$R = \sigma^2 + \text{bias}^2 + \mathcal{O}(n^{-4/5}),$$

just like in univariate smoothing, because the additive model uses univariate smoothing. Because linear models are a special case of additive models, the bias for additive models is less than or equal to the bias for linear models.

They are simple to compute and to interpret (for example, we can examine the effect of each X_j on Y individually while holding all the other variables fixed) so additive models are often a good starting point.

The main downside: By restricting the estimate to be additive, we miss potential interactions between variables. However, like in linear regression, we can manually add *interactions terms* like $r_{ij}(X_i, X_j)$ and $r_{ijk}(X_i, X_j, X_k)$, etc. to the model if we find it to be appropriate.

Backfitting

Reading: Shalizi section 8.2

Given pairs $(\mathbf{X}_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, with each $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip}) \in \mathbb{R}^p$, the additive model becomes

$$Y_i = \beta_0 + r_1(X_{i1}) + \cdots + r_p(X_{ip}) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

subject to the same identifiability assumptions $\mathbb{E}[Y_i] = \beta_0$, and $\mathbb{E}[r_j(X_{ij})] = 0$ for $j = 1, \ldots, p$.

There is a simple algorithm, called *backfitting*, for turning any one-dimensional regression smoother into a method for fitting additive models. The intuition comes from rearranging (3). Suppose that we fixed the intercept and all of the underlying regression functions except the jth one at the estimates $\hat{\beta}_0$ and $\hat{r}_k$, $k \neq j$. The model then becomes

Notes:

$$Y_i - \hat{\beta}_0 - \sum_{k \neq j} \hat{r}_k(X_{ik}) = r_j(X_{ij}) + \epsilon_i, \quad i = 1, \ldots, n.$$

On the left are the *partial residuals*, and we regress these on the columns of X.

To estimate r_j, therefore, we can just treat the left-hand side above as the outcome, and regress this outcome on $X_{\cdot j}$, which is what we do in each iteration of the backfitting algorithm. (Finally, there is a post-centering step to preserve the zero mean condition for model identifiability.)

The Backfitting Algorithm

Initialization: set $\hat{\beta}_0 = \overline{Y}$ and set initial guesses for $\hat{r}_1, \ldots, \hat{r}_p$.

Iterate until convergence: for $j = 1, \ldots, p$,

1. Compute the jth partial residual

$$\widetilde{Y}_i^{(j)} = Y_i - \hat{\beta}_0 - \sum_{k \neq j} \hat{r}_k(X_{ik}), \;\; i = 1, \dots, n.$$

2. Regress $\tilde{Y}^{(j)}$ on the n observations of the jth variable, $X_{\cdot j}$, to obtain a new $\hat{r}_j$, using a univariate regression.

3. Center $\hat{r}_j$,

$$\hat{r}_j = \hat{r}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{r}_j(X_{ij}).$$

Note: There is no reason to use the same univariate smoother in every iteration of backfitting; we can, if we think it is appropriate, use different types of smoothers for different variables. A common choice is to use smoothing splines for each variable, where we either specify the degrees of freedom of the fit ahead of time, or choose it by (generalized) cross-validation in each regression.

R Demo 8.1

Examine the `Wage` data from the ISL book.

(a) Fit an additive model of wage on year, age, and education with smoothing splines for year and age, and a step function for education.

(b) In these plots, the function of year looks rather linear. Try an additive model that excludes year (Model 1), an additive model that uses a linear function of year (Model 2), and compare it to the additive model above that uses a spline function of year (Model 3). Use an F-test (for linear smoothers) to determine which of these three models is best.

Notes: We're fitting an additive model,

$$\text{wage} = \beta_0 + r_1(\text{year}) + r_2(\text{age}) + r_3(\text{education}) + \epsilon.$$

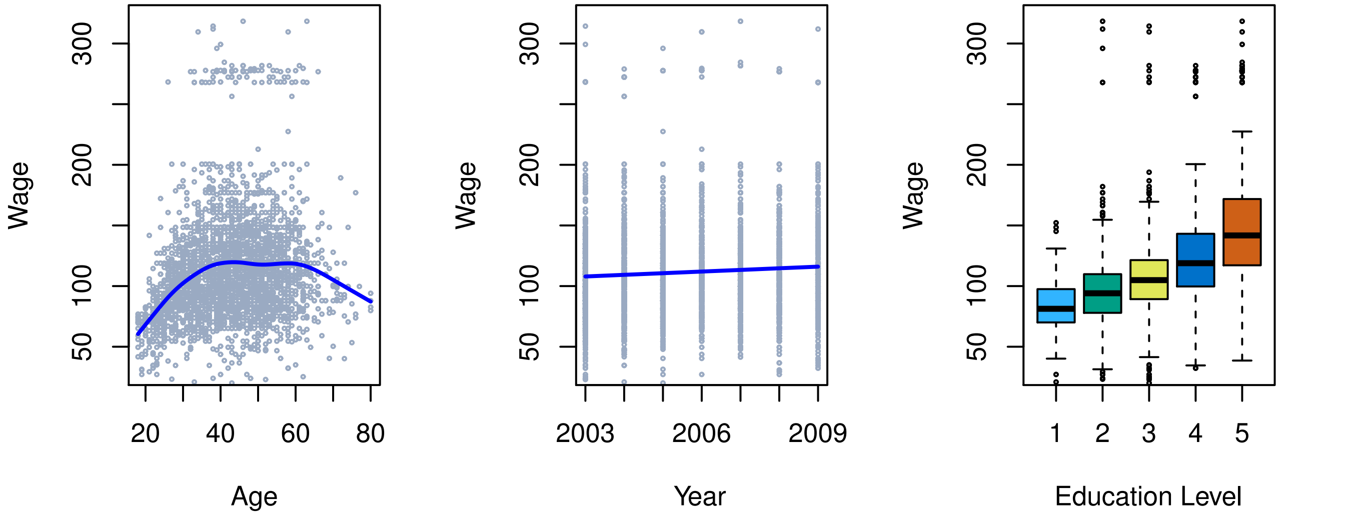

Figure 1: `Wage` data, which contains income survey information for males from the central Atlantic region of the United States. Left: `wage` as a function of `age`. On average, `wage` increases with `age` until about 60 years of age, at which point it begins to decline. Center: `wage` as a function of `year`. There is a slow but steady increase of approximately $10,000 in the average wage between 2003 and 2009. Right: Boxplots displaying `wage` as a function of `education`, with 1 indicating the lowest level (no high school diploma) and 5 the highest level (an advanced graduate degree). On average, `wage` increases with the level of education. [Ref: ISL]

We have three models in part (b). In GAM's notation, they are

M1 wage $\sim s(\text{age}, 5) + \text{education}$, with 10 degrees of freedom $(1 + 5 + 4)$

M2 wage $\sim \text{year} + s(\text{age}, 5) + \text{education}$, with 11 degrees of freedom $(2 + 5 + 4)$

M3 wage $\sim s(\text{year}, 4) + s(\text{age}, 5) + \text{education}$, with 14 degrees of freedom $(1 + 4 + 5 + 4)$.

These models are nested, so we can use the F test: M1 $\subseteq$ M2 $\subseteq$ M3.