# ch07_notes

August 10, 2019

Table of Contents

7 Moving Beyond Linearity
7.1 Polynomial Regression
7.2 Step Functions
7.3 Basis Functions
7.4 Regression Splines
7.4.1 Piecewise Polynomials
7.4.2 Constraints and Splines
7.4.3 The Spline Basis Representation
7.4.4 Choosing the Number and the Locations of the Knots
7.4.5 Comparison to Polynomial Regression
7.5 Smoothing Splines
7.5.1 An Overview of Smoothing Splines
7.5.2 Choosing the Smoothing Parameter $\lambda$
7.6 Local Regression
7.7 Generalized Additive Models
7.7.1 GAMs for Regression Problems
7.7.2 GAMs for Classification Problems
7.8 Footnotes

---

# 1 Moving Beyond Linearity

---

## 1.1 Polynomial Regression

- ***Simple polynomial regression*** is a regression model which is polynomial54 in the feature variable X

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i X^d$$

- The model can be fit as a simple linear regression model with predictors $X_1, \ldots, X_d = X, \ldots X^d$.
- It is rare to take $d \geqslant 4$ because it lead strange curves

**Advantages**

- Interpretability
- More flexibility than linear regression, can better model non-linear relationships

**Disadvantages**

- Greater flexibility can lead to overfitting (can be mitigating by keeping $d$ low)
- Imposes global structure on target function (as does linear regression)

### 1.2 Step Functions

- Step functions model the target function as locally constant by converting the continuous variable $X$ into an ***ordered categorical variable***.as follows
  - Choose $K$ points $c_1, \ldots, c_K \in [\min(X), \max(X)]$
  - Define $K + 1$ "dummy" variables

$$\begin{aligned} C_0(X) &= I(X < c_1) \\ C_i(X) &= I(c_i \leqslant X < c_{i+1}) \qquad 1 \leqslant i \leqslant K-1 \\ C_K(X) &= I(c_K \leqslant X) \end{aligned}$$

  - Fit a linear regression model to the predictors $C_1, \ldots, C_K$55

**Advantages**

- Flexibility to model non-linear relationships
- Can model local behavior better than global models (e.g. linear and polynomial regression)

**Disadvantages**

- Locally constant assumption is strong, breakpoints in data may not be realized.

### 1.3 Basis Functions

In general, we can fit a regression model

$$Y = \beta_0 + \sum_{i=1}^{K} b_i(X)$$

where the $b_i(X)$ are called ***basis functions*** 56

**Advantages** Different choices of basis functions are useful for modeling different types of relationships (for example, Fourier basis functions can model periodic behavior).

**Disadvantages**

- As usual, greater flexibility can lead to overfitting
- Some choices of basis functions (i.e. basis functions which are not suited to the assumed true functional relationship) will likely have poor performance.

## 1.4 Regression Splines

***Regression splines*** are a flexible (and common choice of) class of basis functions which extend both polynomial and piecewise constant basis functions.

### 1.4.1 Piecewise Polynomials

***Piecewise polynomials*** fit separate low-degree polynomials over different regions of $X$. The points where the coefficients change are called ***knots***.

**Advantages**

- Flexibility to model non-linear relationships (as with all non-linear methods discussed in this chapter)
- Sensitivity to local behavior (less rigid than global model).

**Disadvantages**

- Overly flexible - each piece has independent degrees of freedom
- Can have unnatural breaks at knots without appropriate constraints
- Possibility of overfitting (as with all non-linear methods discussed in this chapter)

### 1.4.2 Constraints and Splines

- To remedy overflexibility of piecewise polynomials, we can impose constraints at the knots, e.g. continuity, differentiability of various orders (smoothness).
- A ***spline*** is a piecewise degree $d$ polynomial that has continuous derivatives up to order $d-1$ at each knot (hence everywhere).

**Advantages**

- Same advantages to piecewise polynomials, while improving on the disadvantages

**Disadvantages**

- Overfitting
- Poor match to the true relationship

### 1.4.3 The Spline Basis Representation

- Regression splines can be modeled using an appropriate basis, of which there are many choices.
- For example, we can model a $d$ degree spline with $K$ knots using ***truncated power basis***

$$b_1(X), \dots, b_{K+d}(X) = x, \dots, x^d, h(X, \xi_1), \dots, h(X, \xi_K)$$

  where $\xi_i$ is the $i-th$ knot and

$$h(X-\xi_i) = \begin{cases} (X-\xi_i)^d & X > \xi_i \\ 0 & X \leqslant \xi_i \end{cases}$$

  is the ***truncated power function*** of degree $d$.

**Advantages** Ibid.

**Disadvantages** Beyond those mentioned above, splines can have a high variance near $\min(X), \max(X)$ (this can be overcome by using ***natural splines*** which impose boundary constraints, i.e constraints on the form of the model on $[\min(X), \xi_1]$, $[\max(X), \xi_K]$ (e.g. linearity)

### 1.4.4 Choosing the Number and the Locations of the Knots

- In practice, we place knots in uniform fashion, e.g. by specifying the desired degrees of freedom and using software to place the knots at uniform quantiles of the data.
- The desired degrees of freedom (hence number of knots) can be obtained using cross-validation.

### 1.4.5 Comparison to Polynomial Regression

Often gives superior results to polynomial regression – the latter must use higher degrees (imposing global structure) while the former can increase the number of knots while leaving the degree fixed (sensitivity to local behavior) as well as varying the density of knots (i.e. placing more where the response varies rapidly, less where it is more stable)

## 1.5 Smoothing Splines

### 1.5.1 An Overview of Smoothing Splines

- A ***smoothing spline*** 57 is a function

$$\hat{g}_\lambda = \underset{g}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 \, dt$$

where $\lambda = 0$ is a tuning parameter58 - $\lambda$ controls the bias-variance tradeoff. $\lambda = 0$ corresponds to the ***interpolation spline*** which fits all the data points exactly and will be thus woefull overfit. In the limit $\lambda \to \infty$, $\hat{g}_\lambda$ approaches the least squares line - It can be show that the function $\hat{g}_\lambda$ is a piecewise cubic polynomial with knots at the unique $x_i$ and continuous first and second derivatives at the knots 59

### 1.5.2 Choosing the Smoothing Parameter $\lambda$

- The parameter $\lambda$ controls the ***effective degrees of freedom*** $df_\lambda$. As $\lambda$ goes from $0 $ to $\infty$, $df_\lambda$ goes from $n$ to 2.
- The effective degress of freedom is defined to be

  $$df_\lambda = \operatorname{trace}(S_\lambda)$$

  where $S_\lambda$ is the matrix such that $\hat{\mathbf{g}}_\lambda = S_\lambda \mathbf{y}$ where $\hat{\mathbf{g}}$ is the vector of fitted values.
- $\lambda$ can be chosen by cross-validation. LOOCV is particularly efficient to compute 60

$$RSS_{cv}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left( \frac{y_i - \hat{g}_\lambda(x_i)}{1 - tr(S_\lambda)} \right)^2$$

**Advantages**

- Flexibility/nonlinearity
- As a shrinkage method, effective degrees of freedom are reduced, helping to balance bias-variance tradeoff and avoid overfitting.

**Disadvantages**

- As usual, flexibility can lead to overfitting

## 1.6 Local Regression

- Computes the fit at a target point by regressing on nearby training observations
- Is ***memory-based*** - all the training data is necessary for computing a prediction
- In multiple linear regression, ***variable coefficient models*** fit global regression to some variables and local to others

**Algorithm: $K$-nearest neighbors regression** Fix the parameter61 $1 \leqslant k \leqslant n$. For each $X_{=}x_0$: 1. Get the neighborhood $N_{i0} = \{k \text{ closest } x_i\}$. 2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point $x_i$ such that such that - each point outside $x_i \notin N_{i0}$ has $K_{i0}(x_i) = 0$. - the furthest point $x_i \in N_{i0}$ has weight zero - the closest point $x_i \in N_{i0}$ has the highest weight. 3. Fit a weighted least squares regression

$$(\hat{\beta_0}, \hat{\beta_1}) = \sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

4. Predict $\hat{f}(x_0) = \hat{\beta_0} + \hat{\beta_1} x_0$.

## 1.7 Generalized Additive Models

A ***Generalized additive model*** is a model which is a sum of nonlinear functions of the individual predictors.

### 1.7.1 GAMs for Regression Problems

- A GAM for regression 62 is a model

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

where the functions $f_j$ are smooth non-linear functions.

- GAMs can be used to combine methods from this chapter – one can fit different nonlinear functions $f_j$ to the predictors $X_j$ 63
- Standard software can fit GAMs with smoothing splines via ***backfitting***

**Advantages**

- Nonlinearity hence flexibility
- Automatically introduces nonlinearity - obviates the need to experiment with different nonlinear transformations
- Interpretability/inference - the $f_j$ allow to consider the effect of each feature $X_j$ independently of the others.
- Smoothness of individual $f_j$ can be summarized via degrees of freedom.
- Represents a nice compromise betwee linear and fully non-parametric models (see ğ8).

**Disadvantages**

- Usual disadvantages of nonlinearity
- Doesn't allow for interactions between features (this can be overcome by including nonlinear functios of the interaction terms $f(X_j, X_k)$
- The additive constraint is strong, restricts flexibility.

### 1.7.2 GAMs for Classification Problems

GAMs can be used for classification. For example, a GAM for logistic regression is

$$\log\left(\frac{p_k(X)}{1 - p_k(X)}\right) = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

where $p_k(X) = \Pr(Y = k \mid X)$.

---

## 1.8 Footnotes

54. In statistical literature, polynomial regression is sometimes referred to as linear regression. This is because the model is linear in the population parameters $\beta_i$.

55. The variable $C_0(X)$ accounts for an intercept. Alternatively fit a linear model to $C_0, \dots, C_K$ with no intercept.

56. Such a model amounts to the assumption that the target function lives in a finite-dimensional subspace of the vector space of all functions $f : X \to Y$.

57. The function $g$ is not guaranteed to be smooth in the sense of infinitely differentiable. The penalty on the second derivative (curvature) penalizes the "roughness" or "wiggliness" of $g$, hence "smoothes out" noise in the data. Other penalties have been used

58. A tuning parameter is also called a hyperparameter

59. Thus $\hat{g}$ is a natural cubic spline with knots at the $x_i$. However, it is not the spline one obtains in Section 1.4.3. It is a "shrunken" version, where $\lambda$ controls the shrinkage.

60. Compare to a similar formula in ğ5.1.2

61. Our description of the algorithm deviates a bit from the book, but it's equivalent.

62. “Additive” because we are summing the $f_i$. “Generalized” because it generalizes from the linear functions $\beta_j X_j$ in ordinary linear regression.

63. It’s not hard to see that (with the exception of local regression), all the models discussed in this chapter can be seen as special cases of GAM.