Ch 7.1-7.2: Polynomial regression and Step Functions

Lecture 14 - CMSE 381

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

March 13, 2024

Announcements

Last time:

- PLS
- High dimensions

This lecture:

- 7.1 Polynomial regression
- 7.2 Step functions

Section 1

Last time

High-Dimensional Data

Low-Dimensions

$$n \gg p$$

- Low here means p is low, or at least small relative to n
- Can do all the stuff we've talked about so far

High-Dimensions

$$n \ll p$$

- Issues show up even if $p \geq n$
- Classical approaches not appropriate since lots of overfitting

What to do about it?

Be less flexible....

Key points

- regularization or shrinkage plays a key role in high-dimensional problems,
- appropriate tuning parameter selection is crucial for good predictive performance, and
- the test error tends to increase as the dimensionality of the problem increases, unless the additional features are truly associated with the response.

- Curse of dimensionality
- Report results on an independent test set, or cross-validation errors.

Section 2

Polynomial Regression

Polynomial regression

Replace linear model

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$$

with

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \varepsilon_i$$

- Can learn with linear models by passing in predictors x_i^ℓ
- Tend to not go higher than degree 3 or 4 because makes overly flexible

Coding bit

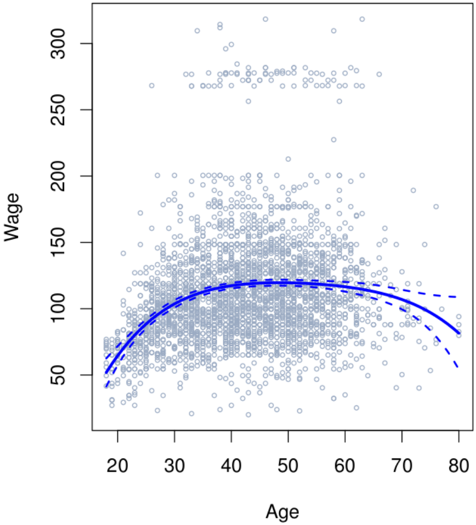
$$\texttt{wage} = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{age}^2 + \cdots + \beta_p \texttt{age}^p + \varepsilon.$$

- My code learned:

$$-184.1542 + 21.24552 * age + -0.56386 * age^2 + 0.00681 * age^3 + -3e - 05 * age^4$$

- Equivalent figure from the book on the next page

Example with wage data



- Plot of wage vs age for men in central Atlantic region of the US
- Dark line is degree 4 polynomial
- Have variance for each coefficient
- Assume you have a model $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \cdots + \hat{\beta}_d x_0^d$
- Can use that to estimate the pointwise variance $\mathrm{Var}(\hat{f}(x_0))$
- Draw 2 std deviations away from the line, 95% confidence interval

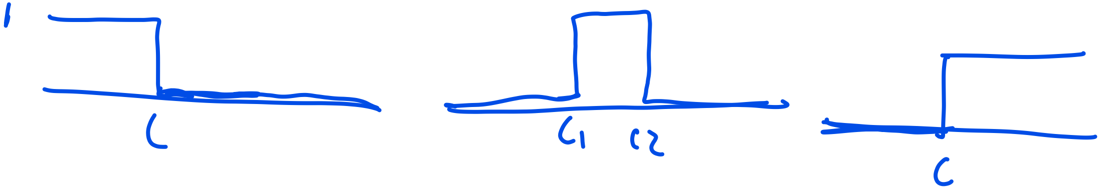
Section 3

Step function

Step functions

$$\mathrm{I}(X < c) \qquad \mathrm{I}(c_1 \leq X < c_2) \qquad \mathrm{I}(c \leq X)$$

- Draw each of the functions above



More on step function setup

because $y = f(x) + \varepsilon$

prediction $\Longleftrightarrow$ function approximation.

Goal of prediction is to find $\hat{f} \approx f$

$$
\begin{cases}
C_0(X) = I(X < c_1), & (-\infty, c_1) \\
C_1(X) = I(c_1 \le X < c_2), & [c_1, c_2) \\
C_2(X) = I(c_2 \le X < c_3), & [c_2, c_3) \\
\quad\vdots & \\
C_{K-1}(X) = I(c_{K-1} \le X < c_K), & \\
C_K(X) = I(c_K \le X), &
\end{cases}
$$

- Choose values $c_1, \cdots, c_K$
- Allow to learn models that don't have global structure
- Use indicator functions to break up the range of X into bins, then we can fit a new constant in each bin.
- these are sometimes also called dummy variables
- Note that $\sum_j C_j(X) = 1$ because X is in only one interval

way to approximate an arbitrary f(x) by step functions

Example

Given knots $c_1 = 3$, $c_2 = 5$, $c_3 = 7$, determine the entries in the columns for $C_i(X)$ in the below matrix.

x	$C_0(X)$	$C_1(X)$	$C_2(X)$	$C_3(X)$
	$(-\infty, 3)$	$[3\ 5)$	$[5, 7)$	$[7, \infty)$
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	1	0	0
5	0	0	1	0

x	$C_0(X)$	$C_1(X)$	$C_2(X)$	$C_3(X)$
6	0	0	1	0
7	0	0	0	1
8	0	0	0	1
9	0	0	0	1
10	0	0	0	1

Draw the function

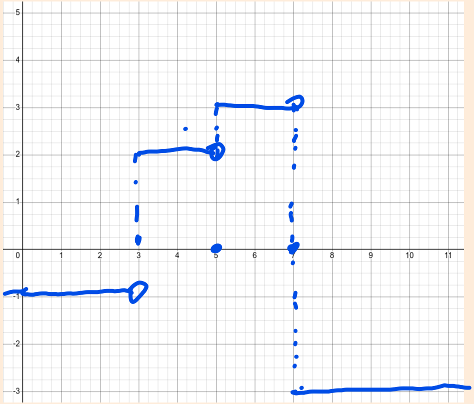
My code doing regression on the step function input returned the function.

$$F(X) = -1 + 3C_1(X) + 4C_2(X) - 2C_3(X).$$

Fill in the table of values, then draw this function below.

X	$F(X)$
1	-1
2	
3	2
4	
5	3

X	$F(X)$
6	
7	-3
8	
9	
10	



Step function: Learned model

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \varepsilon_i$$

$$\underset{\substack{\beta_0, \beta_j \\ j=1,\dots K}}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 C_1 + \cdots + \beta_K C_K \right) \right)^2$$

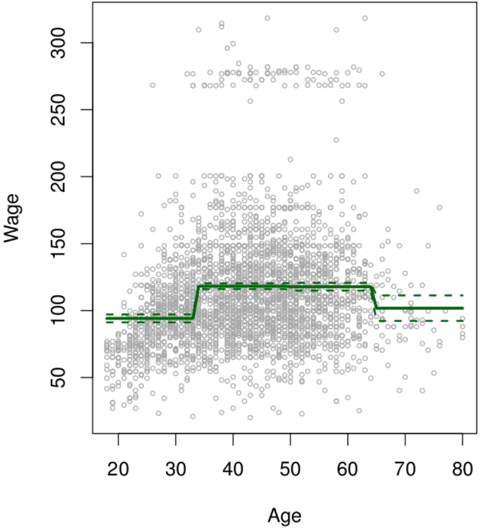
- Note above we don't learn a coeff for C_0 since like qualitative variable version, we can figure out the value from the others.
- If $X < c_1$, all predictors are 0 so β_0 is mean value of Y for $X < c_1$
- Then response for $X \in [c_j, c_{j+1})$ is $\beta_0 + \beta_j$, so β_j is the avg increase in response for X in the interval relative to $X < c_1$

Coding bit

Back to the `wage` data set

Coding with step functions

Step function example



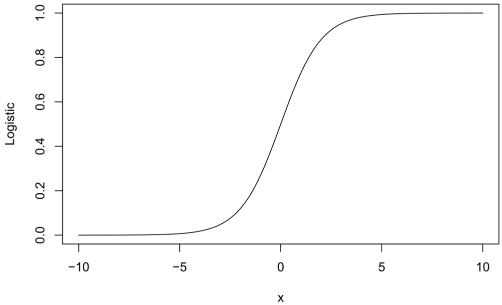
- Learned peak where the high earners are

Section 4

Classification versions

Remember logisitic regression?

$$y = \frac{e^x}{1 + e^x}$$



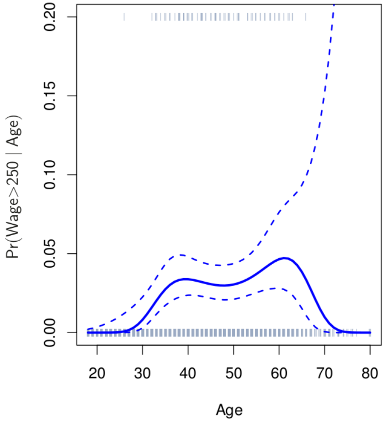
$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Multiple features:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Classification version: Polynomial regression

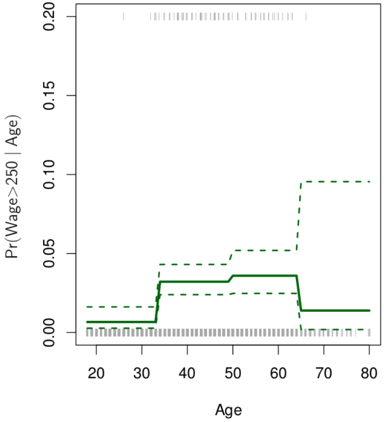
$$\Pr(y_i > 250 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d)}$$



- Note in previous fig that there is a distinct subcluster of high earners making more than \$250K
- Build a logistic regression model as above
- Note that the 95% confidence interval gets very wide on the right side
- Large sample size $n = 3,000$ but small number (79) of high earners makes for high variance in coeffs and wide confidence intervals

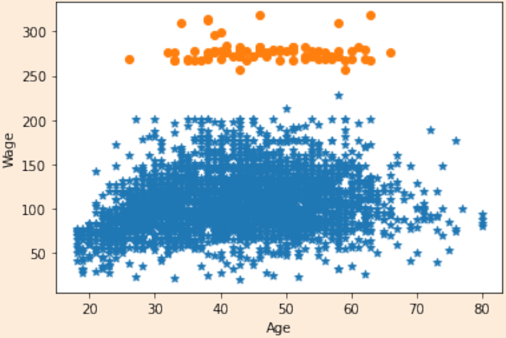
Classification version: Step functions

$$\Pr(y_i > 250 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i))}$$



- Again, learns that middle bit
- Still big confidence interval on the right, but likely because there's just less data over there

Coding bit: classification version



Just talk through on projector, there's nothing in there for them to code

A few more comments on step functions

- Gives the chance to break up the domain, avoid forcing global structure
- Need to make decisions about the c_i. A bit arbitrary unless your data has natural breakpoints.
- Popular in biostats and epidemiology

Section 5

Basis functions

Basis Functions Setup

Polynomial and piecewise-constant regression models are special cases of a *basis function* approach.

$$y_i = \beta_0 + \beta_1 \underline{b_1(x_i)} + \beta_2 b_2(x_i) + \cdots + \beta_K b_K(x_i) + \varepsilon_i$$

- Pick a collection of basis functions $b_1(X), \cdots, b_K(X)$
- Use least squares to figure out the constants
- Explain the b_i's for polynomial and stepwise functions

- Lots of possible options for these
  - Some examples are wavelets or Fourier series
  - Next section is we'll talk about regression splines