

# Kernel Regression

Advanced Methods for Data Analysis (36-402/36-608)

Spring 2014

## 1 Linear smoothers and kernels

- Recall our basic setup: we are given i.i.d. samples  $(x_i, y_i)$ ,  $i = 1, \dots, n$  from the model

$$y_i = r(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

and our goal is to estimate  $r$  with some function  $\hat{r}$ . Assume for now that each  $x_i \in \mathbb{R}$  (i.e., the predictors are 1-dimensional)

- We talked about consider  $\hat{r}$  in the class of linear smoothers, so that

$$\hat{r}(x) = \sum_{i=1}^n w(x, x_i) \cdot y_i \quad (1)$$

for some choice of weights  $w(x, x_i)$ . Indeed, both linear regression and  $k$ -nearest-neighbors are special cases of this.

- Here we will examine another important linear smoother, called *kernel smoothing* or *kernel regression*. We start by defining a kernel function  $K : \mathbb{R} \rightarrow \mathbb{R}$ , satisfying

$$\int_{-\infty}^{\infty} K(t) dt = 1, \quad K(x) = K(-x)$$

- Three common examples are the box kernel:

$$K(x) = \begin{cases} 1/2 & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2),$$

and the Epanechnikov kernel:

$$K(x) = \begin{cases} 3/4(1 - x^2) & \text{if } |x| \leq 1 \\ 0 & \text{else} \end{cases}$$

- Given a choice of kernel  $K$ , and a bandwidth  $h$ , kernel regression is defined by taking

$$w(x, x_i) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right)}$$

in the linear smoother form (1). In other words, the kernel regression estimator is

$$\hat{r}(x) = \frac{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \cdot y_i}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}$$

- What is this doing? This is a weighted average of  $y_i$  values. Think about laying down a Gaussian kernel around a specific query point  $x$ , and evaluating its height at each  $x_i$  in order to determine the weight associated with  $y_i$
- Because these weights are smoothly varying with  $x$ , the kernel regression estimator  $\hat{r}(x)$  itself is also smoothly varying with  $x$ ; compare this to  $k$ -nearest-neighbors regression
- What's in the choice of kernel? Different kernels can give different results. But many of the common kernels tend to produce similar estimators; e.g., Gaussian vs. Epanechnikov, there's not a huge difference
- A much bigger difference comes from choosing different bandwidth values  $h$ . What's the tradeoff present when we vary  $h$ ? Hint: as we've mentioned before, you should always keep these two quantities in mind ...

## 2 Bias and variance of kernels

- At a fixed query point  $x$ , recall our fundamental decomposition

$$\mathbb{E}[\text{TestErr}(\hat{r}(x))] = \mathbb{E}[(Y - \hat{r}(x))^2 | X = x] = \sigma^2 + \text{Bias}(\hat{r}(x))^2 + \text{Var}(\hat{r}(x)).$$

So what is the bias and variance of the kernel regression estimator?

- Fortunately, these can actually roughly be worked out theoretically, under some smoothness assumptions on  $r$  (and other assumptions). We can show that

$$\text{Bias}(\hat{r}(x))^2 = (\mathbb{E}[\hat{r}(x)] - r(x))^2 \leq C_1 h^2$$

and

$$\text{Var}(\hat{r}(x)) \leq \frac{C_2}{nh},$$

for some constants  $C_1$  and  $C_2$ . Does this make sense? What happens to the bias and variance as  $h$  shrinks? As  $h$  grows?

- This means that

$$\mathbb{E}[\text{TestErr}(\hat{r}(x))] = \sigma^2 + C_1 h^2 + \frac{C_2}{nh}.$$

We can find the best bandwidth  $h$ , i.e., the one minimizing test error, by differentiating and setting equal to 0: this yields

$$h = \frac{C_2}{2C_1 n^{1/3}}.$$

Is this a realistic choice for the bandwidth? Problem is that we don't know  $C_1$  and  $C_2$ ! (And even if we did, it may not be a good idea to use this ... why?)

## 3 Practical considerations, multiple dimensions

- In practice, we tend to select  $h$  by, you guessed it, cross-validation
- Kernels can actually suffer bad bias at the boundaries ... why? Think of the asymmetry of the weights

- In multiple dimensions, say, each  $x_i \in \mathbb{R}^p$ , we can easily use kernels, we just replace  $x_i - x$  in the kernel argument by  $\|x_i - x\|_2$ , so that the multivariate kernel regression estimator is

$$\hat{r}(x) = \frac{\sum_{i=1}^n K\left(\frac{\|x_i - x\|_2}{h}\right) \cdot y_i}{\sum_{i=1}^n K\left(\frac{\|x_i - x\|_2}{h}\right)}$$

- The same calculations as those that went into producing the bias and variance bounds above can be done in this multivariate case, showing that

$$\text{Bias}(\hat{r}(x))^2 \leq \tilde{C}_1 h^2$$

and

$$\text{Var}(\hat{r}(x)) \leq \frac{\tilde{C}_2}{nh^p}.$$

Why is the variance so strongly affected now by the dimension  $p$ ? What is the optimal  $h$ , now?

- A little later we'll see an alternative extension to higher dimensions that doesn't nearly suffer the same variance; this is called an additive model

# Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder