

K-NN

↳ FIRST IMPROVEMENT: we give more weight to the observations that are closer to point of interest

↳ we think that closer observations have y_i that are more similar to the one we are interested in than further observations

$$w_i = K_\lambda\left(\frac{x - x_0}{\lambda}\right)$$

$$K_\lambda(x, x_0) = \frac{D(1 - |x - x_0|)}{\lambda}$$

$$\hat{f}_{KNN}(x_0) = \frac{1}{K} \sum_{i \in N_K(x_0)} y_i$$

$$\hat{f}_{LR}(x_0) = \frac{\sum_{i=1}^N K_\lambda(x_i, x_0) y_i}{\sum_{i=1}^N K_\lambda(x_i, x_0)}$$

$$w_i = \frac{K_\lambda(x_i; x_0)}{\sum_{i=1}^N K_\lambda(x_i; x_0)}$$

λ is the tuning parameter (smoothing parameter or bandwidth)

Typical kernels:

normal

$$\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \left(\frac{x_i - x_0}{\lambda}\right)^2\right\}$$



Epanechnikov

$$\frac{3}{4} \left[1 - \left(\frac{x_i - x_0}{\lambda}\right)^2\right]$$

defined in $[-1; 1]$

biquadratic

$$\frac{15}{16} \left[1 - \left(\frac{x_i - x_0}{\lambda}\right)^2\right]^2$$

defined in $[-1; 1]$

tricubic

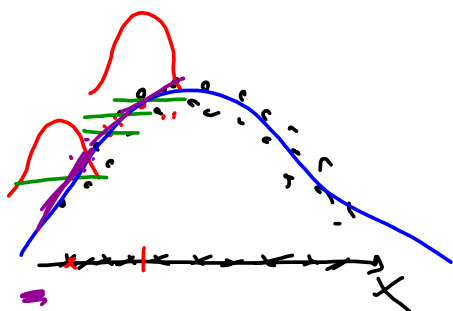
$$\frac{70}{81} \left[1 - \left|\frac{x_i - x_0}{\lambda}\right|^3\right]^3$$

defined in $[-1; 1]$

↳ the choice of the different kernels does not have a huge impact, there are many empirical studies that show that

The really important choice is the value of λ :

- two options:
- 1) choose it in a way that a specific number of observations enter in the window
 - 2) fixed



SECOND IMPROVEMENT: from constant to linear approximation

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

Taylor expansion around x_0

$$f(x) = \underbrace{f(x_0)}_{\beta_0} + \underbrace{f'(x_0)}_{\beta_1} (x - x_0) + \underbrace{o(x - x_0)}_{\text{rest}}$$

order of the error term

We estimate β_0 and β_1 using least squares. In addition, we need to introduce the weights $\rightarrow$ WEIGHTED LEAST SQUARES

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n K_{\lambda}(x_i, x_0) \left[y_i - \beta_0(x) - \beta_1(x_0)x_i \right]^2 \quad (6.7)$$

OLS: $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = (X^T X)^{-1} X^T y$

WLS: $\hat{\beta} = (\hat{\beta}_0(x_0), \hat{\beta}_1(x_0)) = (X^T W X)^{-1} W X^T y$

$$W = \begin{pmatrix} K_{\lambda}(x_1, x_0) & & 0 \\ & \ddots & \\ 0 & & K_{\lambda}(x_n, x_0) \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ x \end{pmatrix} \begin{pmatrix} 1 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix}$$

Choice of the bandwidth

We need to find the best value for λ , as a compromise between:

- smaller λ , smaller bandwidth, estimate is very localised $\Rightarrow$ small bias, large variance $\rightarrow$ too bumpy estimate $\rightarrow$ overfitting
- larger λ , wider bandwidth, observations far from the point of interest are considered $\Rightarrow$ large bias, small variance $\rightarrow$ too flat estimate $\rightarrow$ underfitting

the choice of λ is again based on bias-variance trade-off

In practice, λ is chosen by cross-validation

Formally, under specific conditions,

$$E[\hat{f}_{\lambda}(x)] \approx f(x) + \frac{\lambda^2}{2} \sigma_w^2 f''(x)$$

$$\text{Var}[\hat{f}_{\lambda}(x)] \approx \frac{\sigma_w^2}{n\lambda} \frac{\alpha(w)}{g(x)}$$

with $\sigma_w^2 = \int z^2 w(z) dz$, $\alpha(w) = \int w(z)^2 dz$, $g(x)$ is the density of the random variable from which x_i were sampled

for λ sufficiently small, n sufficiently large

bias = constant $\times \lambda^2$
variance = constant $\times \frac{1}{n\lambda}$

I need a small λ to reduce the bias
I need a large value of λ to reduce the variance

Extension to several dimensions

↳ x_1 and x_2 ...

Our local method to estimate $f(x)$ can be extended to the general case of p covariates (dimensions)

E.g. $p=2$ $y = f(x_1, x_2) + \varepsilon$ $f: \mathbb{R}^2 \rightarrow \mathbb{R}$

Let us consider $i=1, \dots, n$ observations (y_i, x_{1i}, x_{2i}) , $i=1, \dots, n$

We need to find β that minimizes

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^n K_2(x_i, x_0) (y_i - b(x_i)^T \beta(x_0))^2$$

$$K_2(x_i, x_0) = \frac{D(1|x_{1i} - x_0^1|)}{\lambda_1} \frac{D(1|x_{2i} - x_0^2|)}{\lambda_2}$$

↑
↑

[1]
[2]

Although in theory our local (regression) methods work in p dimensions, they are rarely used for $p > 2$

- difficulties to visualize / use in more than 2 dimensions
- they suffer the curse of dimensionality