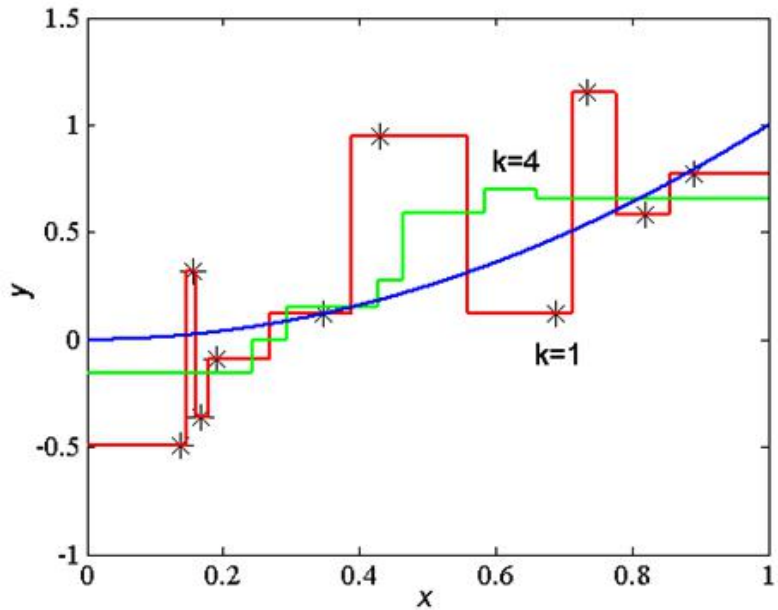
Data Science II
(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

Review: K-nearest neighbors (KNN)

- KNN uses local neighborhood to obtain a prediction
- A distance is needed to compare the similarity
  - Euclidean distance, Manhattan distance
- If the number of dimensions is very high the nearest neighbors can be very far away



Local regression

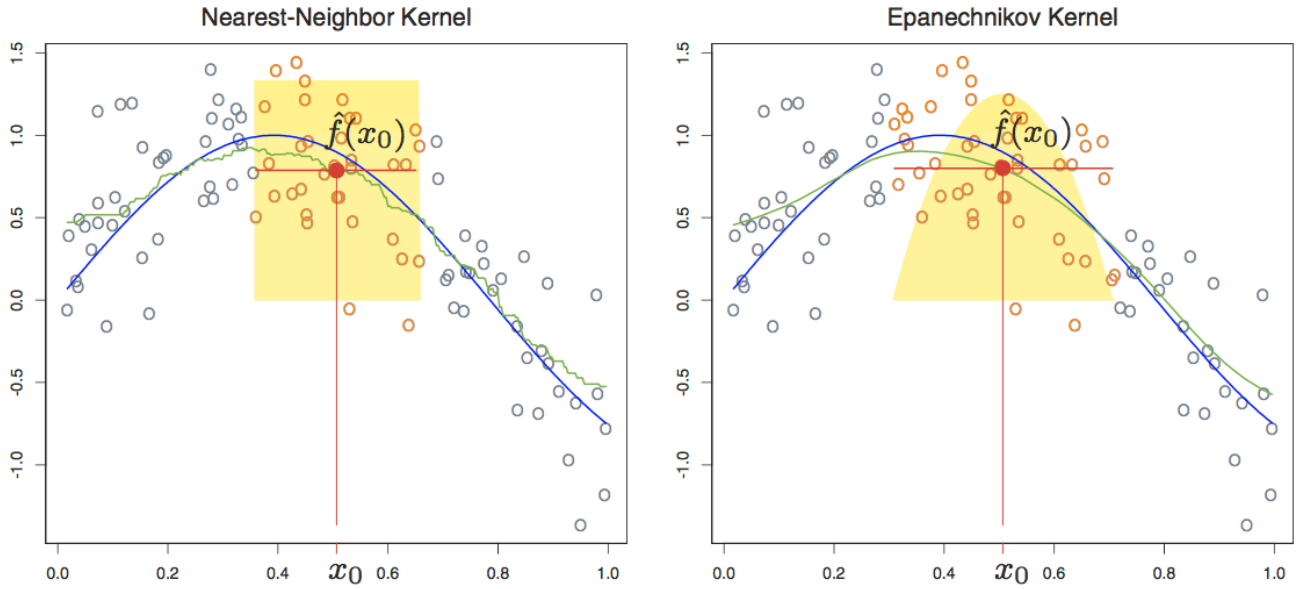
- $\hat{f}(x_0) = \text{Ave}(y \mid x \in N(x_0)) = \sum_{i=1}^{n} w(x_0, x_i) y_i$
- KNN: $w(x, x_i) = I(x_i \in N_K(x))/K$
  - The weight drops to 0 outside $N_K(x)$
- Kernel-based techniques (Nadaraya-Watson estimator)

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

  where $K_\lambda(x_0, x) = D\left(\frac{|x - x_0|}{\lambda}\right)$
- D - kernel function
  - $\int_{-\infty}^{\infty} D(u) du = 1$
  - Usually symmetric around 0
- λ - bandwidth

Kernel function

- Uniform kernel: $D(t) = 0.5I(|t| \leq 1)$
- Epanechnikov kernel: $D(t) = 0.75(1 - t^2)I(|t| \leq 1)$
- Gaussian kernel: $D(t) = \exp(-t^2/2)/\sqrt{2\pi}$

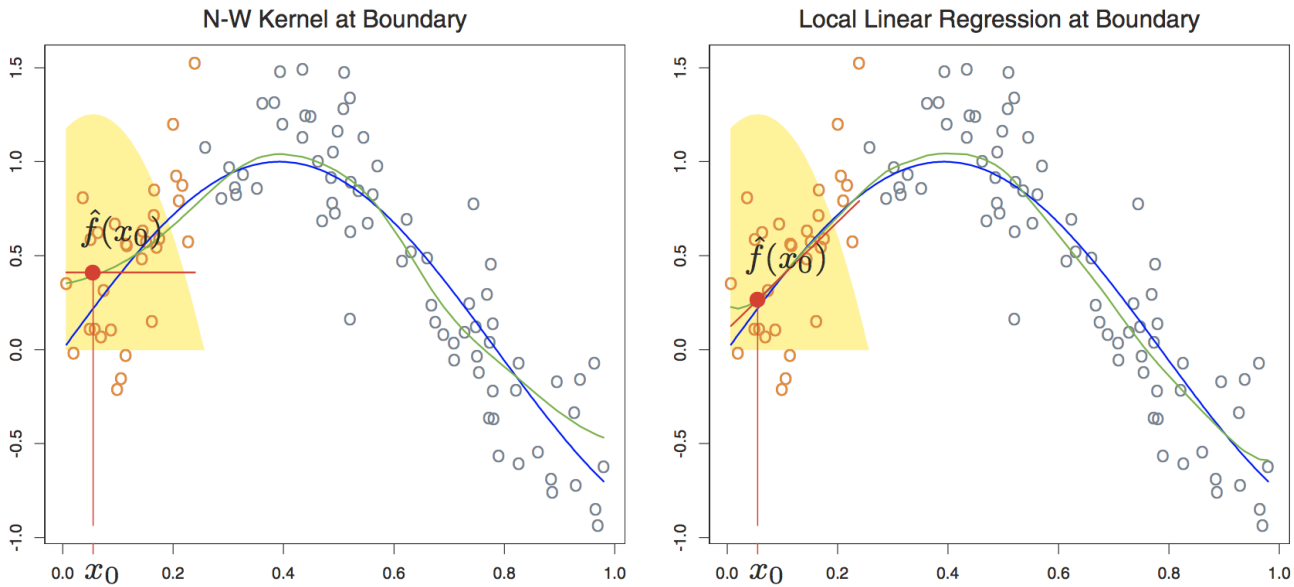


Local linear regression

- Find α_0, β_0 that minimize

$$\sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \alpha_0 - \beta_0 x_i)^2$$

- The estimate is linear in y_i
- Fitted value $\hat{f}(x_0) = \hat{\alpha}_0 + \hat{\beta}_0 x_0$
- Reduce bias near boundary



Generalized additive model (GAM)

- Allows for flexible nonlinearities in several variables, but retains the additive structure of linear models
- $g\{E(Y \mid X)\} = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$
- Identifiability?
- Using the aforementioned methods as building blocks
- Advantages
  - Automatically model non-linear relationships that standard linear regression will miss
  - Can potentially make more accurate predictions

Generalized additive model

- Two packages implement GAM
  - `gam` (textbook): need to specify degree of freedom
  - `mgcv`: simultaneously fit the model and optimize over the smoothing parameters
- Similar syntax but different results
- With the current support from `caret`, you may lose a significant amount of flexibility in `mgcv`

Generalized additive model

- Can mix terms – some linear, some nonlinear
- One can use ANOVA to compare nested models
- Hypothesis testing is often not the purpose of the analysis
- Building a model that accurately estimates the relationship between the outcome and predictors may be a more meaningful goal

Multivariate Adaptive Regression Splines

- Create a piecewise linear model
- Given a cut point c for a predictor, two new features are hinge functions $\{h(x-c), h(c-x)\}$ of the original
- Hinge function $h(x) = x_+$
- The algorithm automatically selects cut points
- Two tuning parameters: the degree of features and the number of terms

Multivariate Adaptive Regression Splines

- R package: `earth` (Enhanced Adaptive Regression Through Hinges)
- Using caret: `method = "earth"`