Nonparametrics and Local Methods: Nearest Neighbor Methods

C.Conlon

February 28, 2023

Applied Econometrics

Bias Variance Decomposition

We can decompose any estimator into two components

$$\underbrace{\mathbb{E}[(y - \hat{f}(x))^2]}_{MSE} = \underbrace{\left(\mathbb{E}[\hat{f}(x) - f(x)]\right)^2}_{Bias^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right)^2\right]}_{Variance}$$

- In general we face a tradeoff between bias and variance.
- More complicated models reduce bias, but at the expense of higher variance.

Bias Variance Decomposition

What minimizes MSE?

$$f(x_i) = \mathbb{E}[Y_i|X_i]$$

- Seems simple enough (but we are back where we started).
- How do we compute the expectation ?
  - OLS uses entire dataset and adds structure $y = x\beta$ to the problem.
  - Can use polynomials in x.
  - k-NN tries to use local information to estimate conditional mean

How about logit?

We write:

$$\mathbb{E}[Y_i|X_i] = Pr(Y_i = 1|X_i) = p(x_i)$$
$$Pr(Y_i = 1|X_i) = \frac{1}{1 + \exp^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots}}$$

Or with the log odds transformation:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

Logit is linear again (in parameters). This is a generalized linear model.

How else might we estimate $\mathbb{E}[Y_i|X_i]$?

Obvious approach:

- With enough data: look at all values for $X_i = x$ and take the mean.
- Easy if X is discrete or doesn't take on too many values (Gender, State/Country).
- Could work if X is continuous but rounded (test scores, years of school, etc.).
- We could cut x_i into distinct bins (like a histogram).

A Fake Data Example

Following the THF textbook example, we can generate some fake data and let:

$$Y = \quad ORANGE \text{ if } Y^* \quad > 0.5$$
$$Y = \quad BLUE \text{ if } Y^* \quad \leq 0.5$$

- Easiest way to recover Y^* is by running OLS on the linear probability model.
- Draws from bivariate normal distribution with uncorrelated components but different means (2 overlapping types)
- Mixture of 10 low variance (nearly point mass) normal distributions where the individual means were drawn from another normal distribution. (10 nearly distinct types).

Linear Probability Model

FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable (*BLUE $= 0$, ORANGE $= 1$*), and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$*. The orange shaded region denotes that part of input space classified as* ORANGE*, while the blue region is classified as* BLUE*.*

Alternative

- Lots of potential alternatives to our decision rule.
- A simple idea is to hold a majority vote of neighboring points

$$Y^* = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

- To avoid including "yourself" in your neighborhood, we often estimate on one sample and validate on another
- How many parameters does this model have: None? One? k?
- Technically it has something like N/k.
- As $N \longrightarrow \infty$ this means we have an infinite number of parameters! (This is a defining characteristic of non-parametrics).

15 Nearest Neighbor

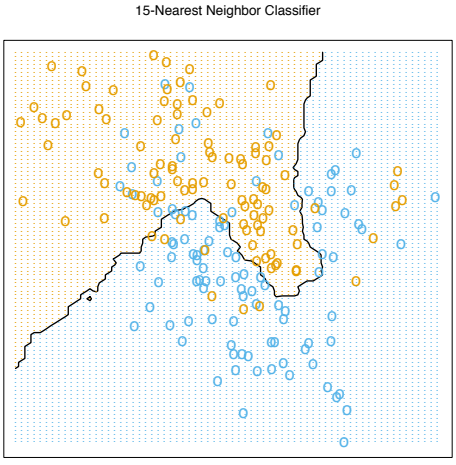

FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$) *and then fit by* 15*-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the* 15*-nearest neighbors.*

Extreme: 1 Nearest Neighbor

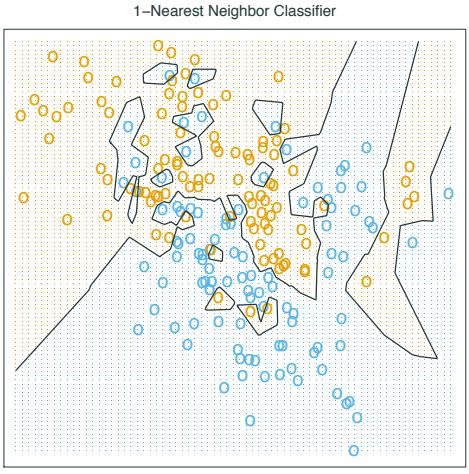

FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$)*, and then predicted by* 1*-nearest-neighbor classification.*

Comparisons

- What would happen if $K \longrightarrow N$?
- The k-NN model is locally constant.
- The k-NN approach tends to be really bumpy which can be undesirable.

What about?

- If we fixed the fact that there are discrete jumps in who is in the neighborhood by smoothly weighting observations and varying those weights instead (Kernels).
- Another drawback of $k - NN$ is that we consider distance in each X dimension on the same scale, perhaps we could rescale the data to improve our "closeness" measure.
- Instead of fitting a constant locally, we fit a linear function locally (Lowess).
- Instead of using a global linear approximation in OLS use a more flexible nonlinear one.
- There is a bias/variance tradeoff. explain.