Non Parametric Models

Dr. Farshid Alizadeh-Shabdiz

Spring 2024

Two Quotes By Famous Statisticians

"Essentially, all models are wrong, but some are useful"

George Box

Reference: An Introduction to Statistical Learning, by Gareth James

SKYHOOK®

Two Quotes By Famous Statisticians

"Essentially, all models are wrong, but some are useful"

George Box

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"

Fred Mosteller and John Tukey, paraphrasing George Box

Reference: An Introduction to Statistical Learning, by Gareth James

SKYHOOK®

Parametric Model

Parametric model explicitly assumes a function to explain the output.

Linear regression is an example of a parametric approach - it assumes a linear functional f(X).

Parametric methods advantages.

- Often easy to fit - a small number of coefficients to estimate.
- Tests of statistical significance can be easily performed.

Parametric methods disadvantages

- By construction makes strong assumptions about f(X).
- If the specified function is far from the truth, and prediction accuracy is our goal, the parametric method won't do well.

SKYHOOK®

Non Parametric Model

- Definition: non-parametric methods do not explicitly assume a parametric form for f(X)

Advantage

- Provide more flexible approach
- Can be applied to regression and classification
- Can be used for any complex relationship

Disadvantage

- Needs more data
- As relationship becomes more complex, more data is needed
- Data is needed everywhere

K Nearest Neighbors

- Example: one of the simplest and best-known non-parametric methods is K-nearest neighbors (KNN)

KNN:

- Given a value for K and a prediction point x0
- first identifies the K training observations that are closest to x0, represented by N0.
- The estimate is the average of all the training responses in N0 for regression and majority of the classes for classification
- $$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$
- The optimal value for K will depend on the bias-variance tradeoff
- A small value for K provides flexible fit, which will have low bias but high variance. Larger values of K provide a smoother and less variable fit; However, the smoothing may cause bias by masking some of the structure in f(X)

SKYHOOK®

Finding the Best K

- Examining the test error rate will help up to choose optimal K value.

SKYHOOK®

When a parametric and When a non-parametric

- the parametric approach will outperform the non- parametric approach if the parametric form that has been selected is close to the true form of f

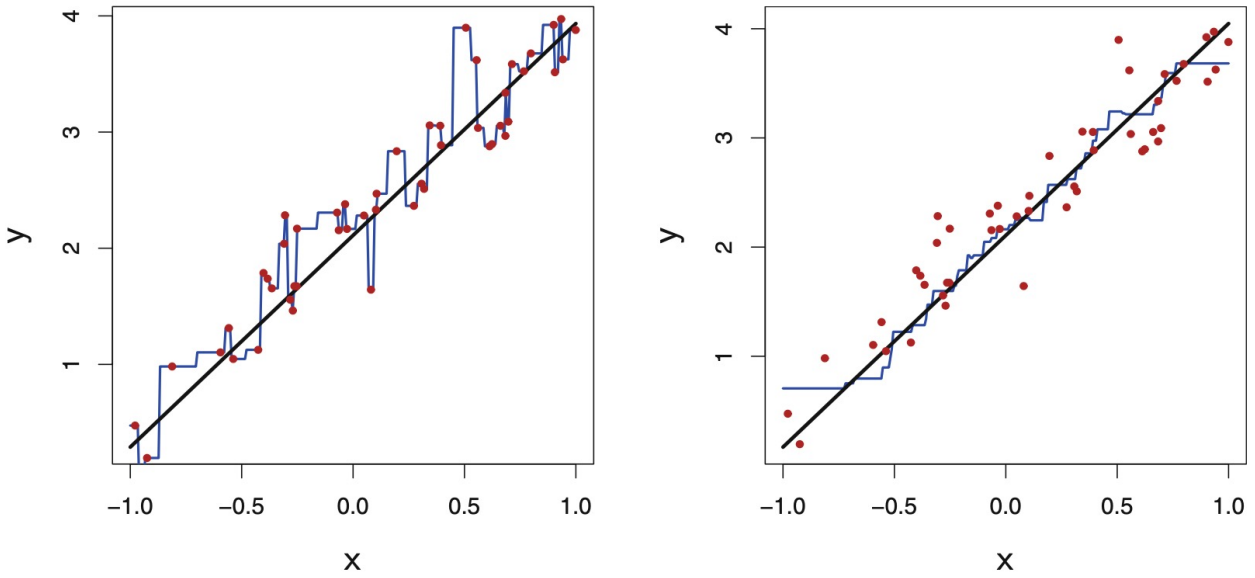

FIGURE 3.17. *Plots of $\hat{f}(X)$ using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line.* Left: *The blue curve corresponds to $K = 1$ and interpolates (i.e. passes directly through) the training data.* Right: *The blue curve corresponds to $K = 9$, and represents a smoother fit.*

Reference: An Introduction to Statistical Learning, by Gareth James

SKYHOOK®

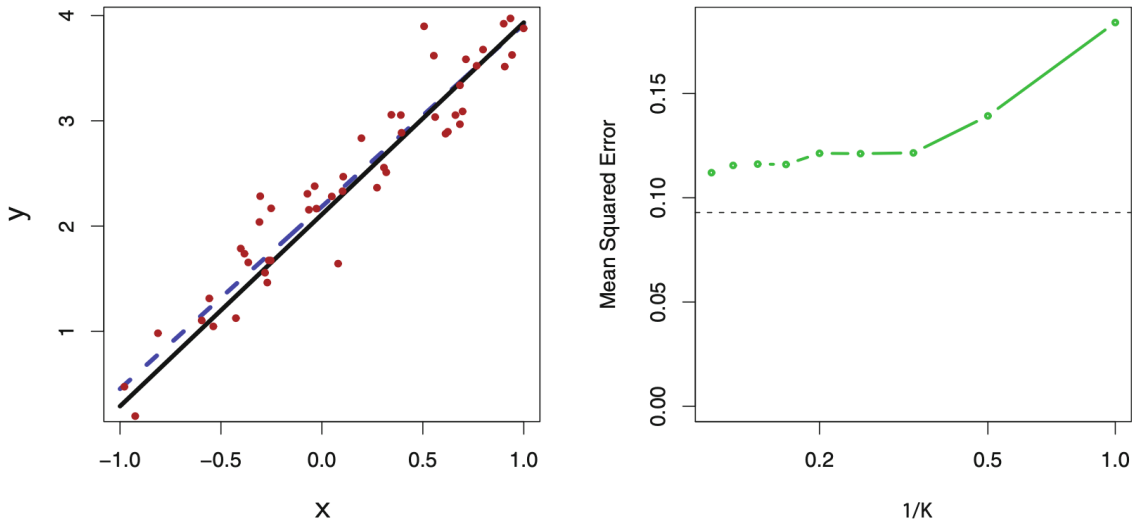

FIGURE 3.18. *The same data set shown in Figure 3.17 is investigated further.* Left: *The blue dashed line is the least squares fit to the data. Since $f(X)$ is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of $f(X)$.* Right: *The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of $1/K$ (on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since $f(X)$ is in fact linear. For KNN regression, the best results occur with a very large value of K, corresponding to a small value of $1/K$.*

Reference: An Introduction to Statistical Learning, by Gareth James

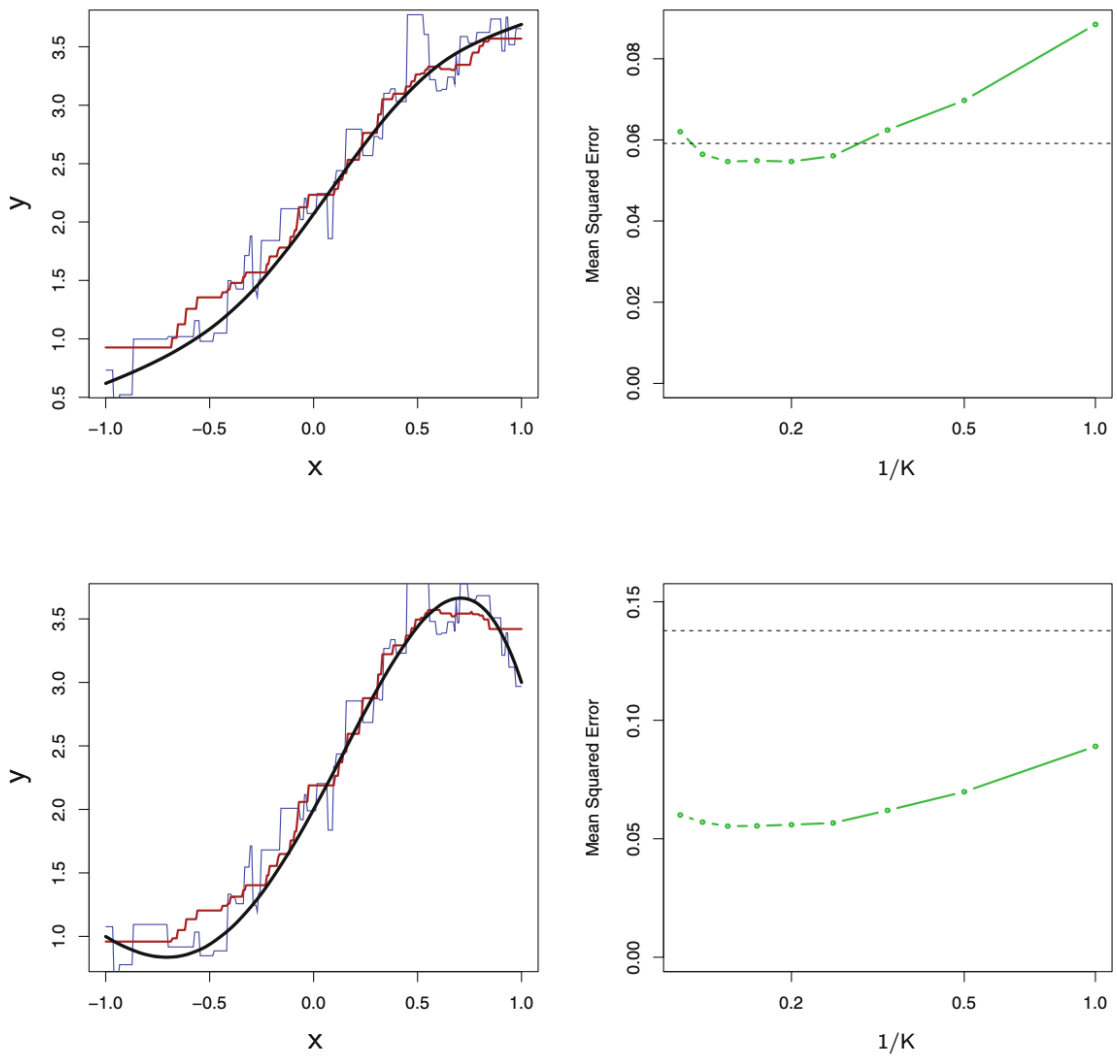

FIGURE 3.19. Top Left: *In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with $K = 1$ (blue) and $K = 9$ (red) are displayed.* Top Right: *For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of $1/K$ (green) are displayed.* Bottom Left and Bottom Right: *As in the top panel, but with a strongly non-linear relationship between X and Y.*

Reference: An Introduction to Statistical Learning, by Gareth James

SKYHOOK®

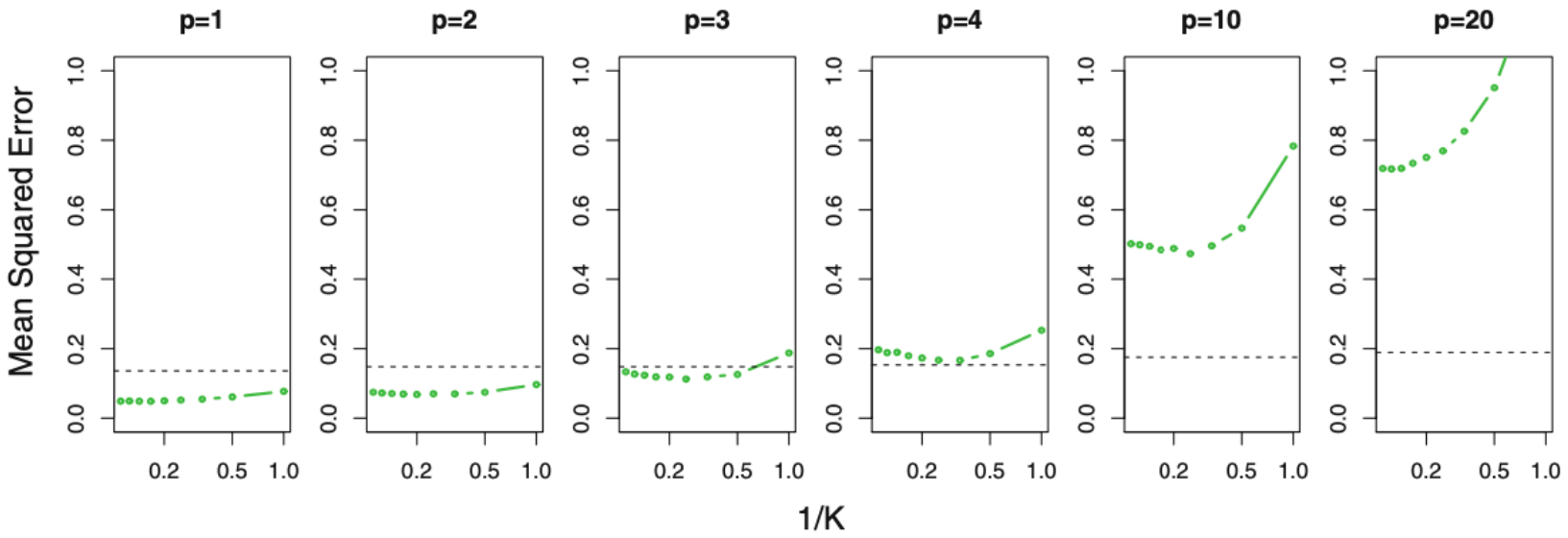

FIGURE 3.20. *Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non–linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.*

Reference: An Introduction to Statistical Learning, by Gareth James

SKYHOOK®

- For high dimensional problems, we might lean toward parametric approach
- Even in low dimension problems, when the MSE of KNN is only slightly worse, we might prefer parametric to KNN from an interpretability standpoint, also it provides a simple model that can be described in terms of just a few coefficients, plus capability of computing p-values.

Summary of KNN Pros and Cons

- None parametric
- No assumption about the data data model
- No assumption about the decision boundary
- Even cover complex boundaries
- Good with large number of observations relative to the number of predictors
- Reduces bias but larger variance of the model itself (smaller error, larger response variation).
- Unlike other classifiers (like Logistic regression) doesn't tell us which parameters are important

SKYHOOK®

K Nearest Neighbor

- Distance measure – Euclidian or something

- Scaling
- Dealing with categorical variables
- How to choose K – cross validation
- Curse of dimensionality
  - Edge issue
    - A random point in a unit square is 0.4% chance of being within 0.001 from the edges
    - But in 10,000 dimensions unit hypercube, this probability is 99.99%
  - Distance between two random points gets larger since there is more space, e.g.
    - in 2D unit square = 0.52
    - In 3D unit cube = 0.66
    - In 1,000,000 dimension = 408.25

SKYHOOK®

KNN and Categorical Variables

In KNN algorithm, categorical variables must get converted into a numerical values so that they can be used in distance calculations.

Common approaches to handle categorical variables in KNN:

1. **Label Encoding:** Convert categorical variables into numerical labels. Each category is assigned a unique integer. For example, categorical variable with values "A", "B", and "C", can have numerical labels like 1, 2, and 3, respectively.
2. **One-Hot Encoding:** Convert categorical variables into binary. For each categorical variable, a new binary column is created, and a 1 is placed in the corresponding column for the category present in the observation, while 0 is placed in all other columns.
3. **Dummy Coding:** Similar to one-hot encoding, but one fewer binary variable is created. This is done to avoid multicollinearity. One category is chosen as a reference, and binary variables are created for all other categories.
4. **Other Distance Metrics:** For categorical variables there are other distance measurements, like Hamming distance, which counts the number of positions at which the corresponding symbols are different.