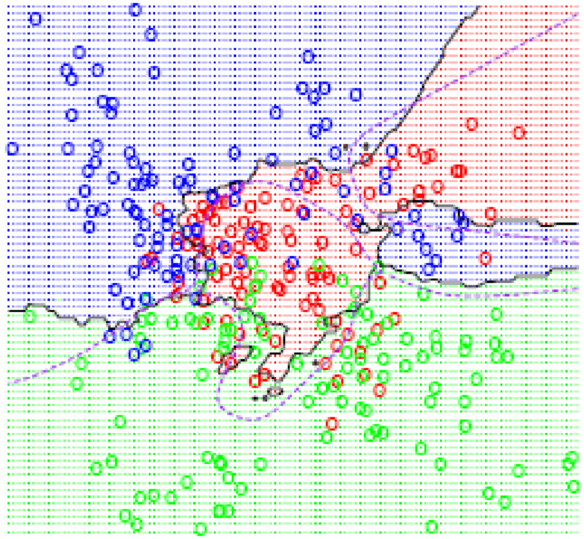
Machine Learning

Lecture 7: K-nearest neighbor regression; classification; KNN classifier; logistic regression

The lectures are mainly offered on white board accompanied by some slides.

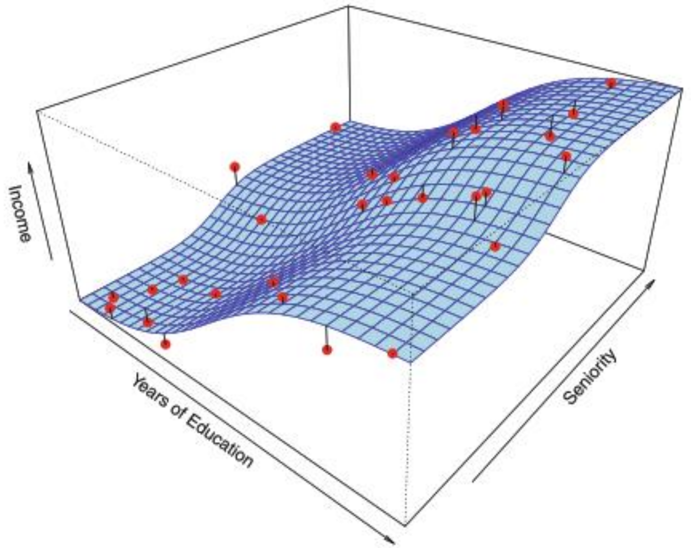


Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

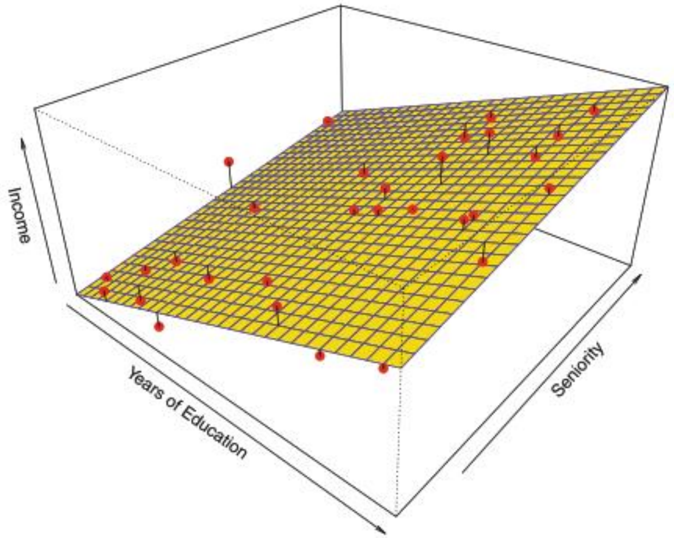
Mehr 21, 1398

Nonparametric methods

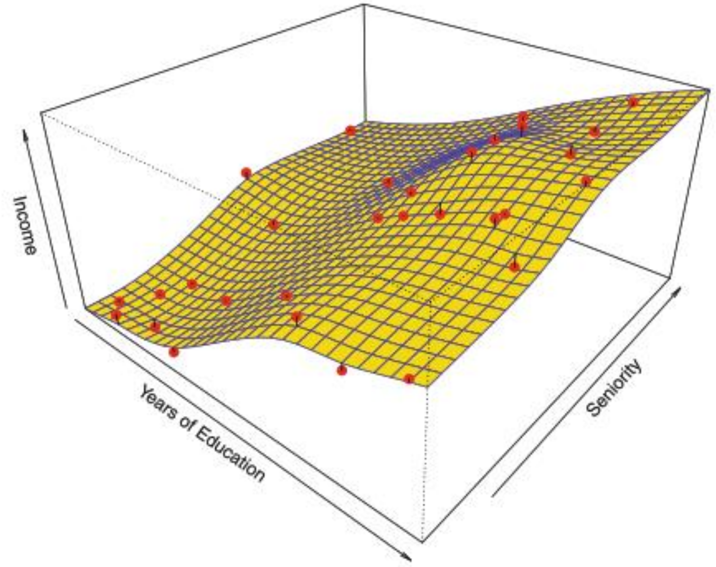
- No assumption about the function form (advantage)
- Try to get as close to data points without being too rough or wiggly
- Disadvantage: need to keep a large number of observations for a good estimate



True underlying relationship



A linear model fit by least squares



A relatively advanced non-parametric method

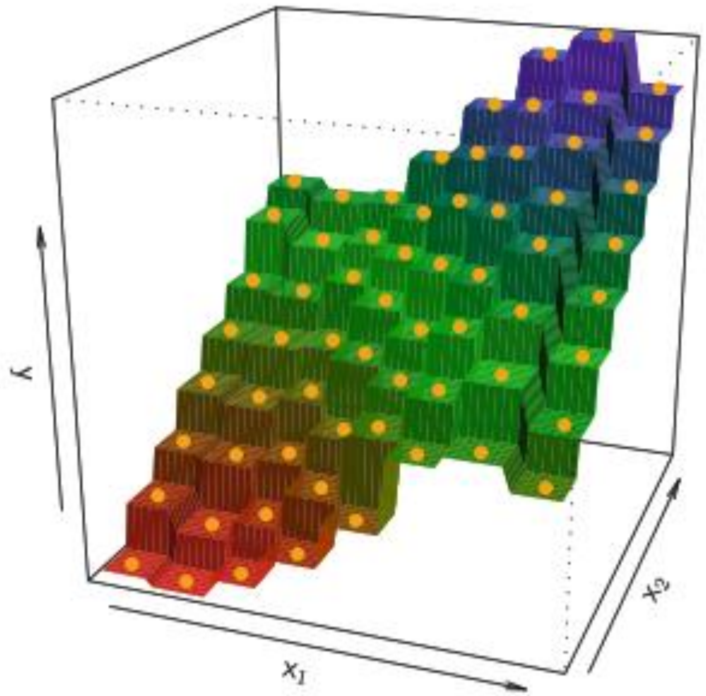
K-nearest neighbors regression (KNN regression)

- Give K and x_0
- N_0: K closest observation to x_0

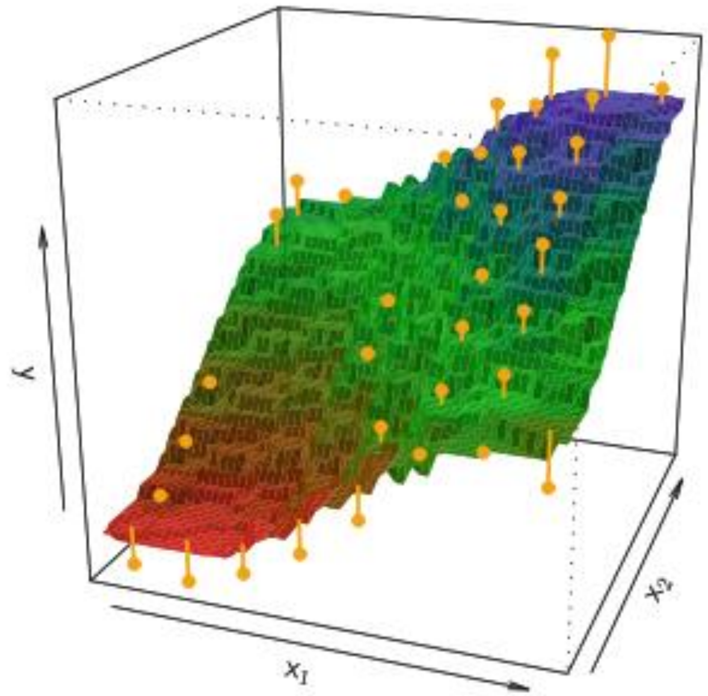
$$\hat{f}(x_0) = \frac{1}{K} \sum_{i \in N_0} y[i]$$

- K = 1 bias=? Variance=?
- K=10 bias=? Variance=?

KNN regression



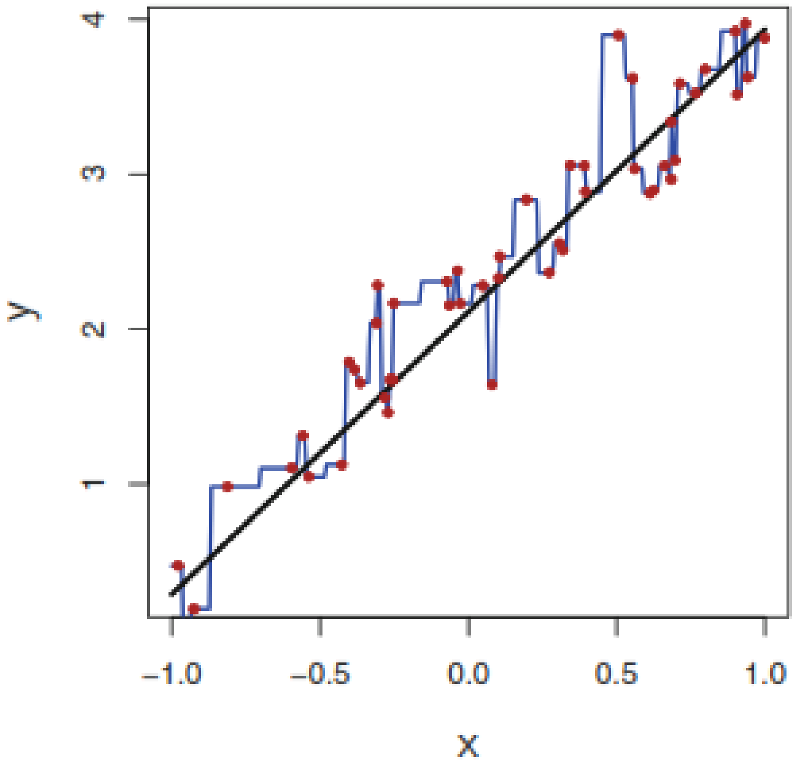
K=1



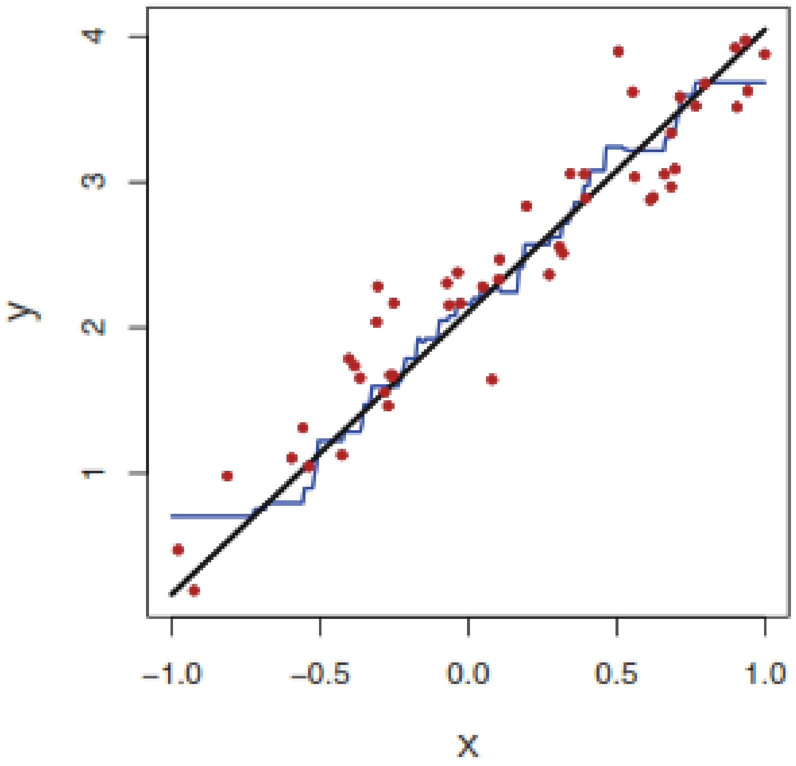
K=9

ISL, Figure 3.16

[Bishop]

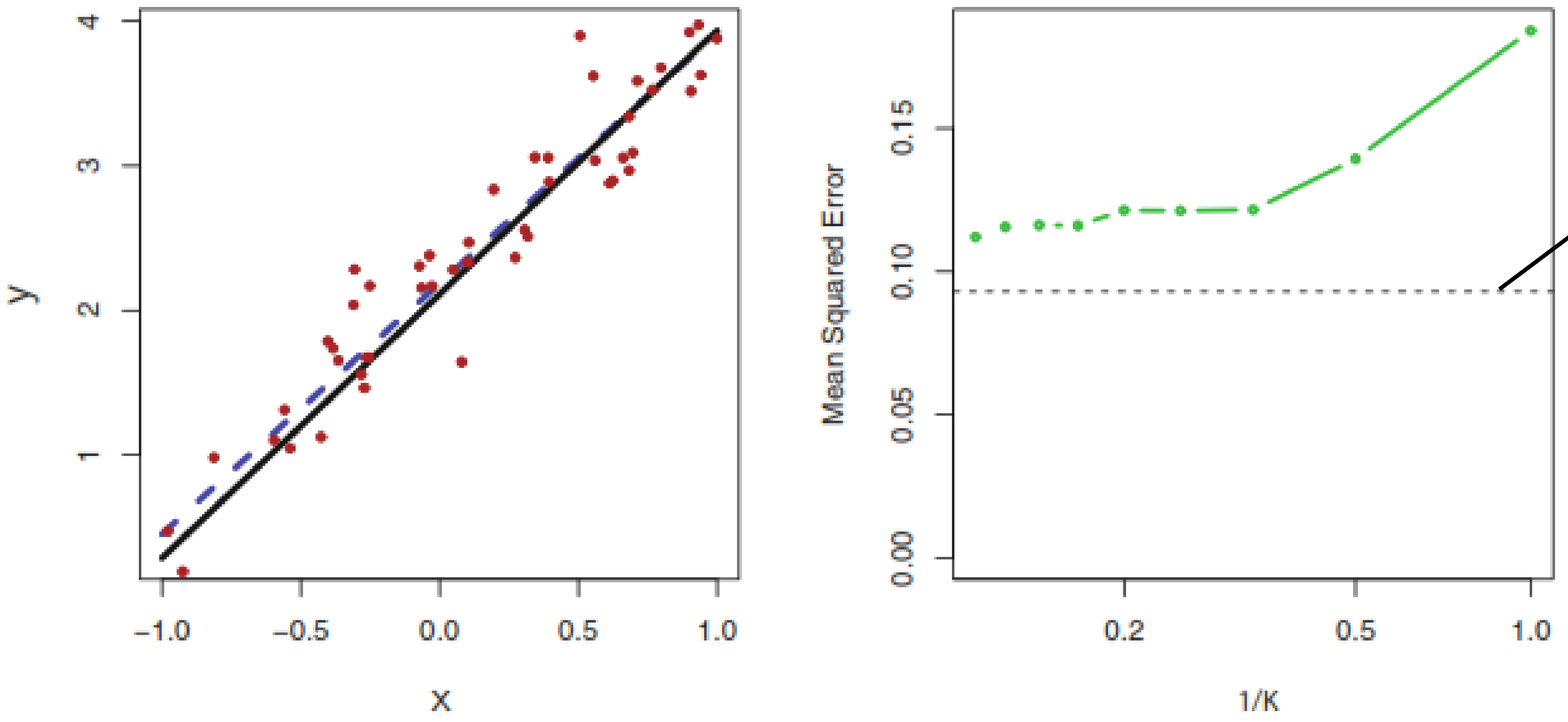


K=1



K=9

ISL, Figure 3.17



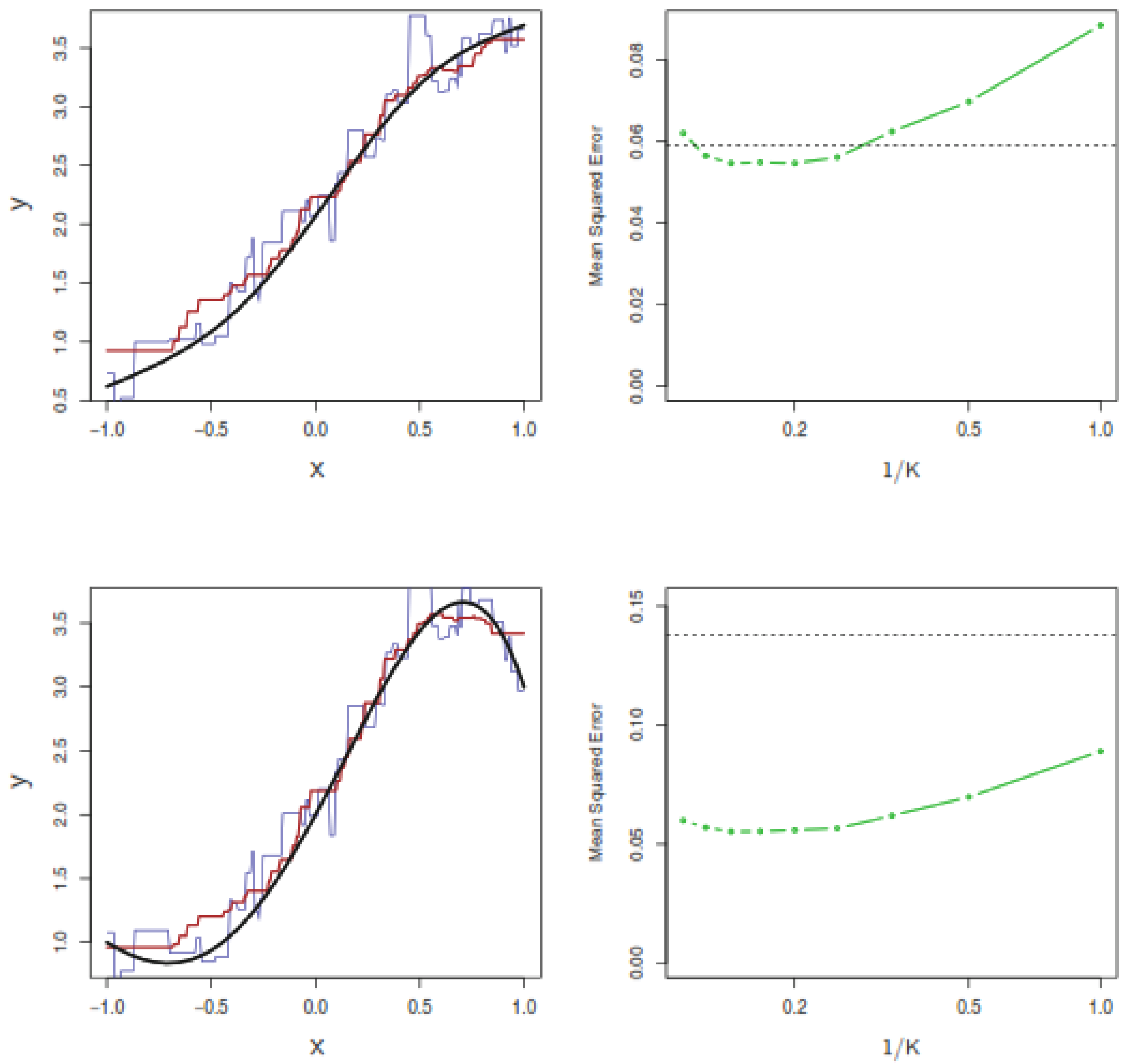
MSE for Least squares fit

Least squares fit

The MSE

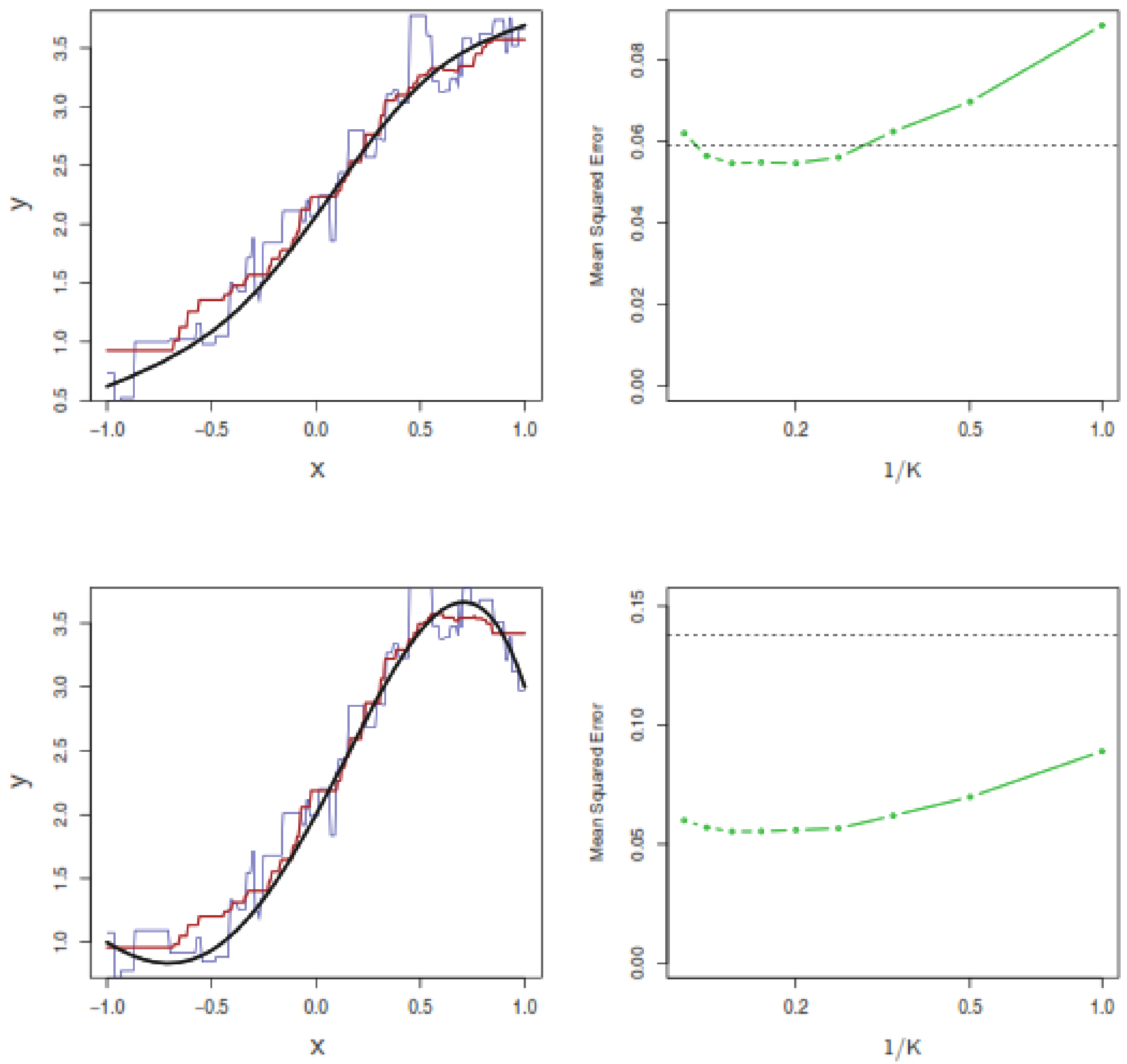
ISL, Figure 3.18

Non-linear relationship between X and Y



ISL, Figure 3.19

Non-linear relationship between X and Y

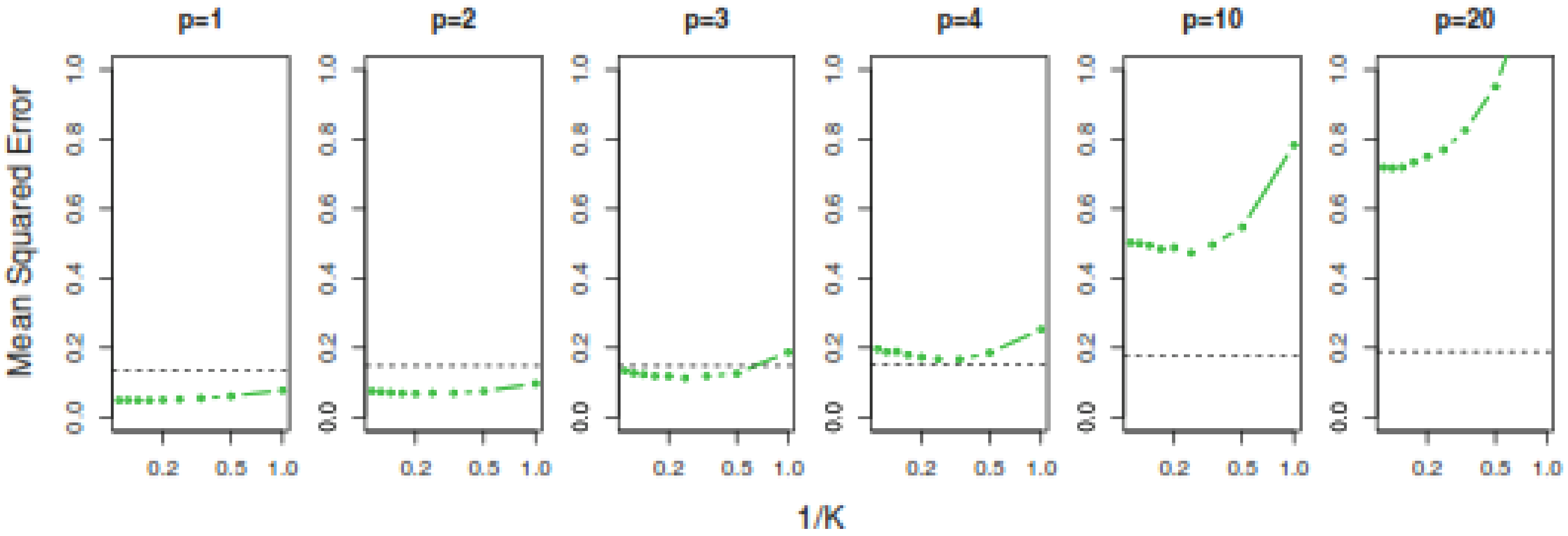


ISL, Figure 3.19

Curse of dimensionality

- K observations that are close to x_0 may be very far away in p-dimensional space when p is large.

- By our definition, Y ONLY depends on the first predictor. The additional predictors are noise.



Classification

Classification

- The response variable is qualitative/categorical.
- $D = \{(x[1], y[1]),\ \ldots, (x[n], y[n])\}$

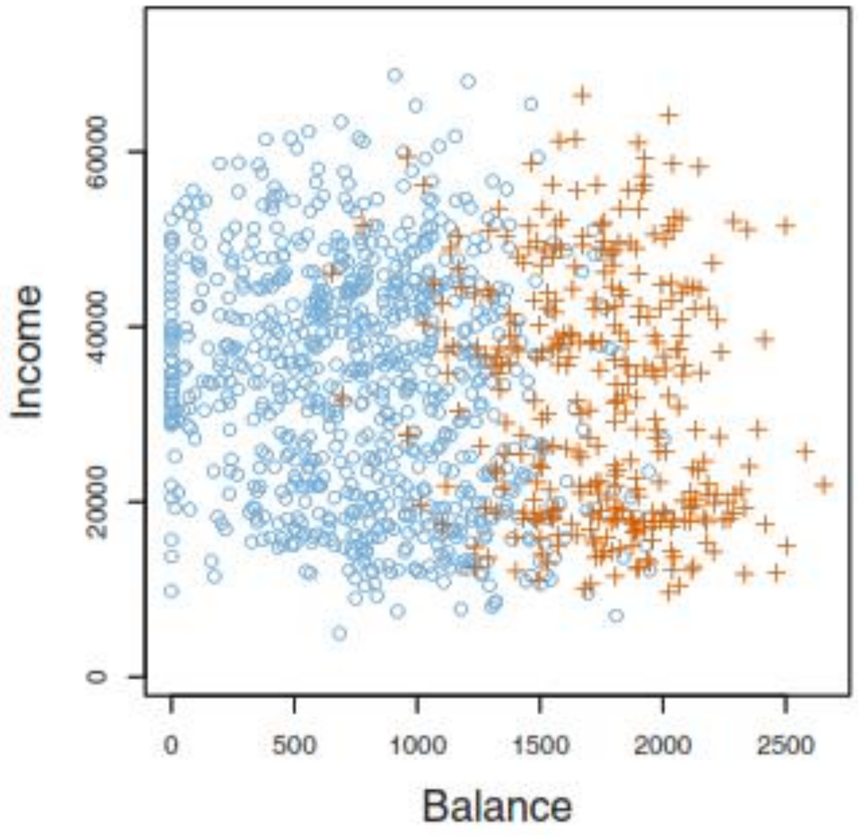
- Examples:
  - Ex1: Input: gene expression data, response variable: origin of cancer (breast, lung, kidney, …)
  - Ex2: to classify an email to spam/non-spam

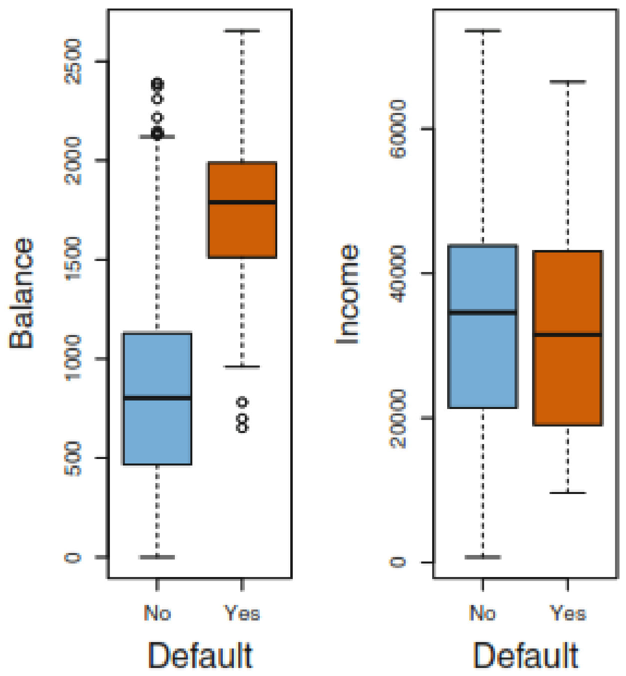
- We define a new loss function:
  - Zero-one loss function: $I(y \neq \hat{y})$ where $\hat{y}$ is an estimate for x

- Training error rate: $\frac{1}{\mathrm{n}} \sum_{i=1}^{n} I(y[i] \neq \hat{y}[i])$

- Test error rate: $E(I(y \neq \hat{y}))$

Example





Response variable:

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

The Bayes classifier

- Assign each data point to the most likely class, given its predictor values

- Formally, the Bayes classifier is defined

$$\hat{y} = \arg\max_{\mathrm{c}} P(Y = c \mid x)$$

- Often not possible in practice! Why?

- The Bayes error rate:
  - The Bayes classifier produces the lowest possible test error rate.
  - The overall error rate is given by

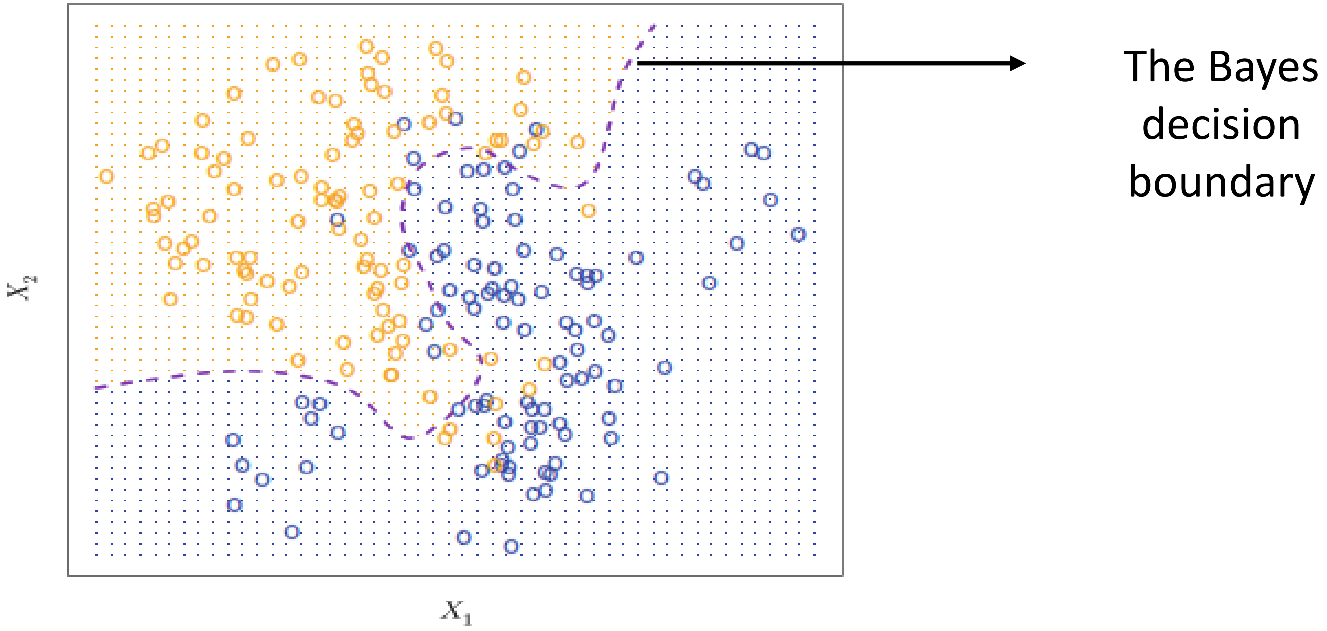
Expectation over all possible values for X

$$1 - E(\max_{c} P(Y = c \mid X)$$

  - Analogous to irreducible error in regression

The Bayes-optimal decision boundary

- Two-class problem: orange and blue



- Given the generating density for each class, we can calculate the boundary exactly.

K-nearest neighbor classifier

- $P(Y \mid X)$ is not known for real data, so computing the Bayes classifier is impossible!

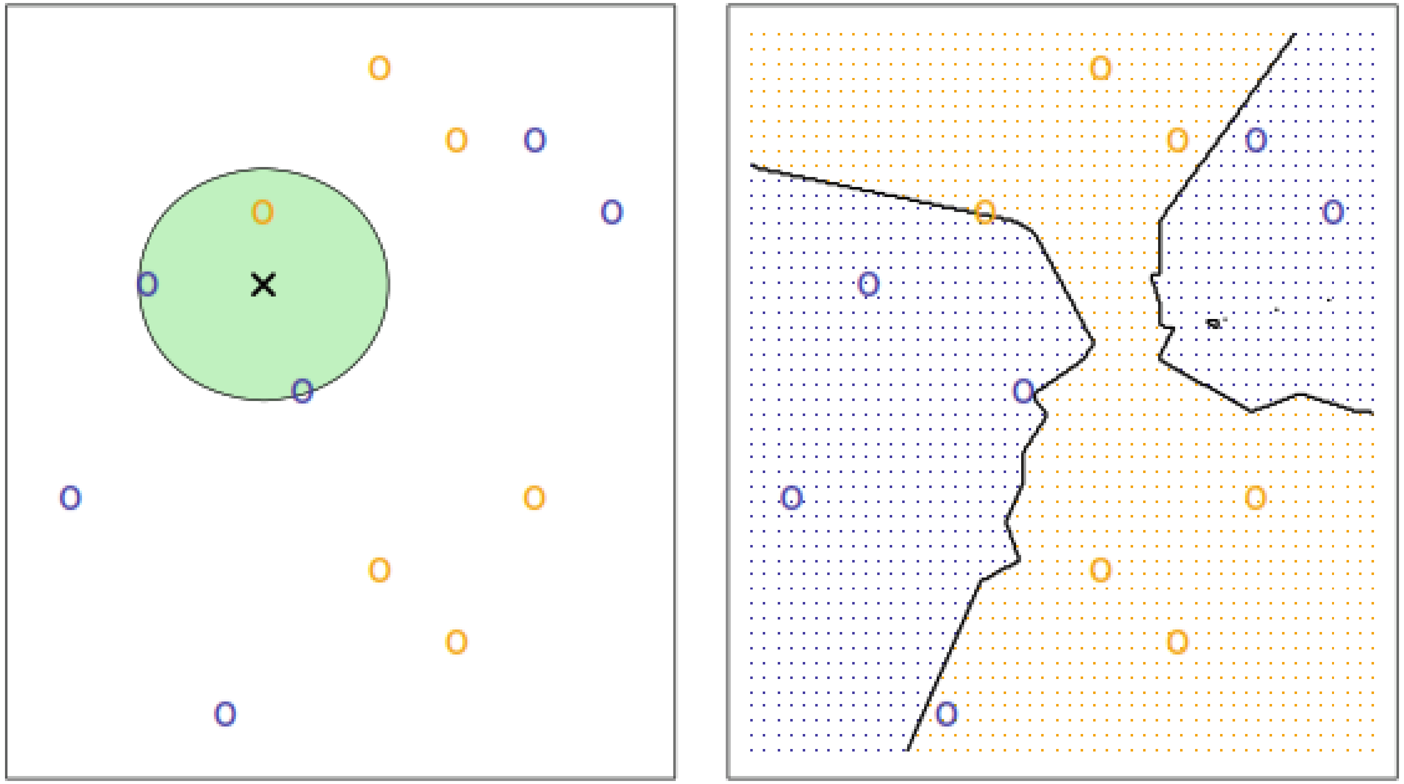
- Many classifiers attempt to estimate $P(\, Y \mid X \,)$

- K-nearest neighbor classifier is such a method:

$$P(\, Y = c \mid x \,) = \frac{1}{K} \sum_{i \in N_0} I(y[i] = c)$$

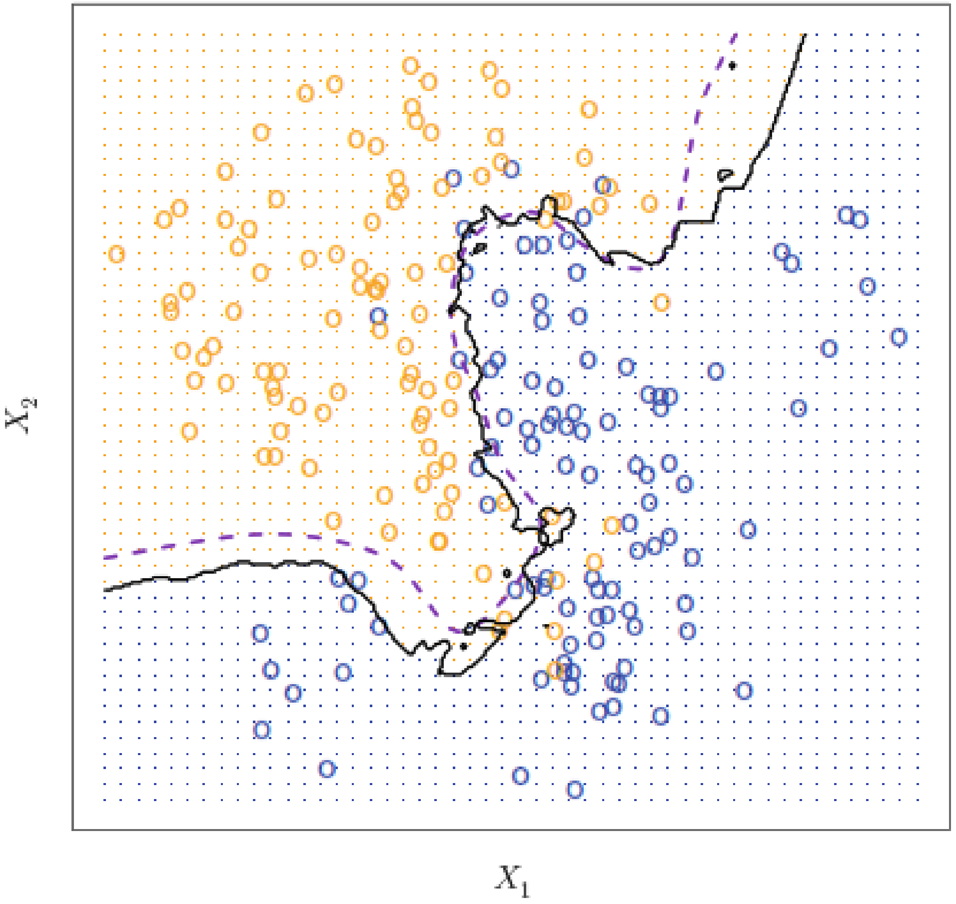
where N_0 is K closest observation to x

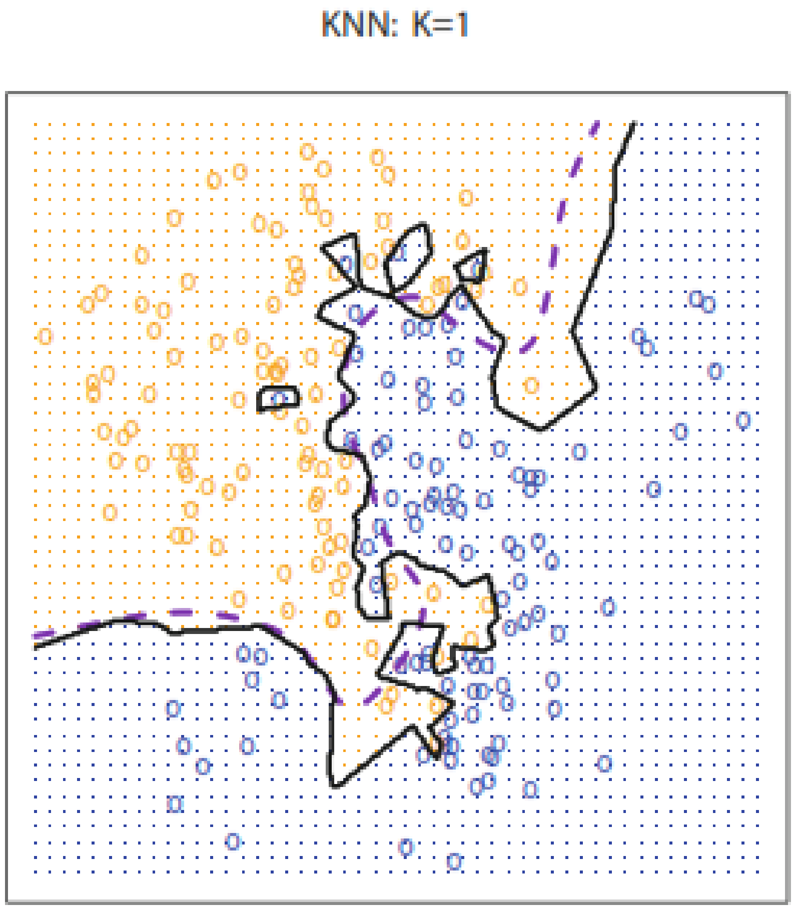
K-nearest neighbors in two dimensions



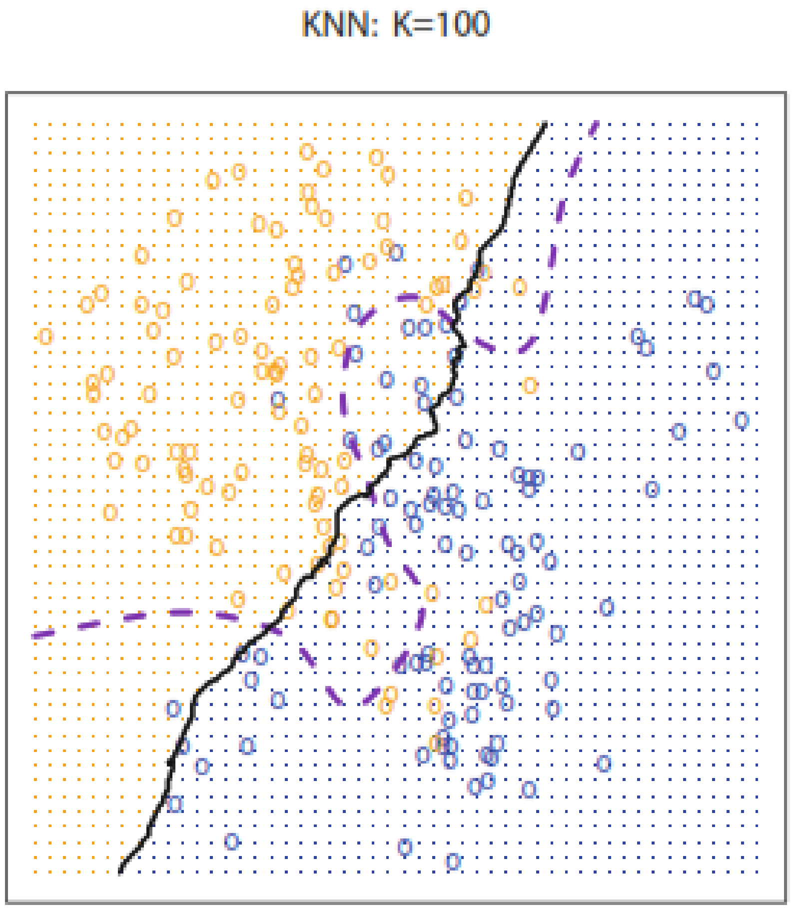
K = 3

K-nearest neighbors: K=10





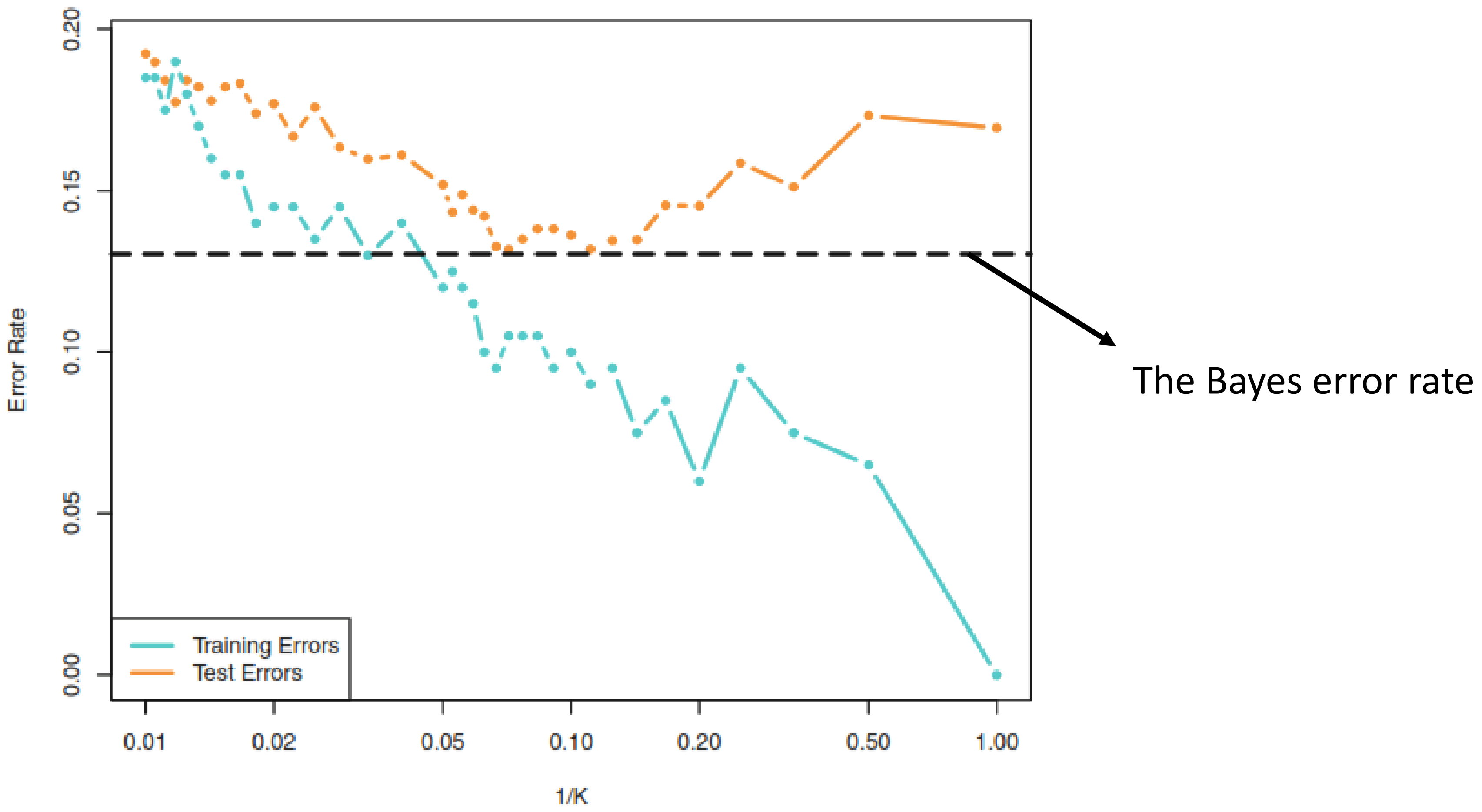
Flexible boundary



In flexible boundary

Training and test error rates

- 200 training observations, 5000 test observations



References and Acknowledgement

- References
  - An Introduction to Statistical Learning, with applications in R, 2013

- Acknowledgement
  - Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani. Few slides are also adjusted from theirs.