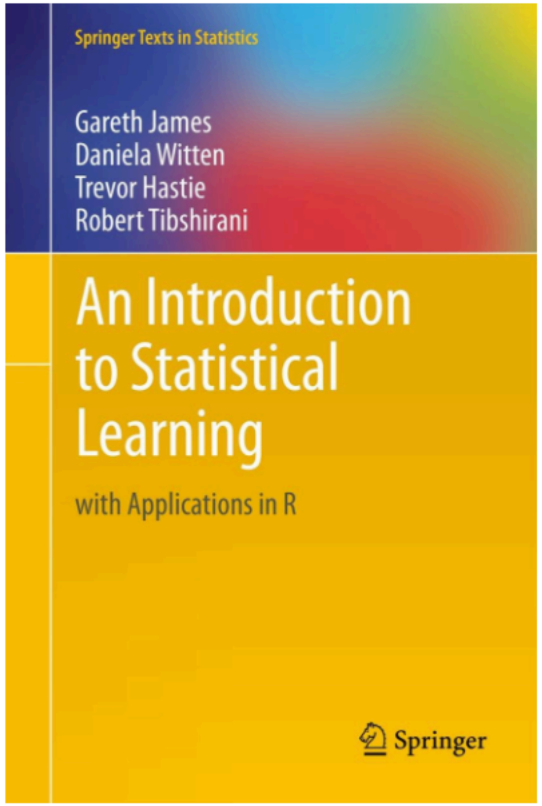
5. K-Nearest Neighborhood

ESC Spring 2018 – Data Mining and Analysis

SeoHyeong Jeong



Textbook:

An Introduction to Statistical Learning

Lecture Slides:

Stanford Stats 202: Data Mining and Analysis

Spring 17’ ESC Statistical Data Analysis

REFERENCE

Reading:

An Introduction to Statistical Learning

chapter 2.2.3 The Classification Setting

Table of Contents

Classification Problems

- Suppose we seek to estimate f on the basis of training observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$, where $y_1, \dots, y_n$ are qualitative.
- Classification task is to build a function $C(X)$ that takes input explanatory variable(s) X and predicts value for Y.
- Our main goals:
  - Building a classifier $C(X)$
  - Assessing the uncertainty

- Classification techniques, or *classifiers*, are used to predict a qualitative response. Most widely used classifiers are:
  - *The Bayes Classifier (not widely used but it's the basic)*
  - *KNN*
  - *logistic regression*
  - *linear discriminant analysis*

Training Error Rate and Test Error Rate

- The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the *training error rate*:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

  where $I(y_i \neq \hat{y}_i)$ is an *indicator variable* that equals 1 if $y_i \neq \hat{y}_i$ and 0 if $y_i = \hat{y}_i$
- Training error rate computes the fraction of incorrect classifications.

- The training error rate is computed based on the data that was used to train the classifier.
- The *test error rate* associated with a set of test observations of the form (x_0, y_0) is given by

$$\text{Ave}\left(I(y_0 \neq \hat{y}_0)\right),$$

where $\hat{y}_0$ is the predicted class label that results from applying the classifier to the test observation with predictor x_0.

- A *good classifier* is one for which the test error is smallest.

The Bayes Classifier

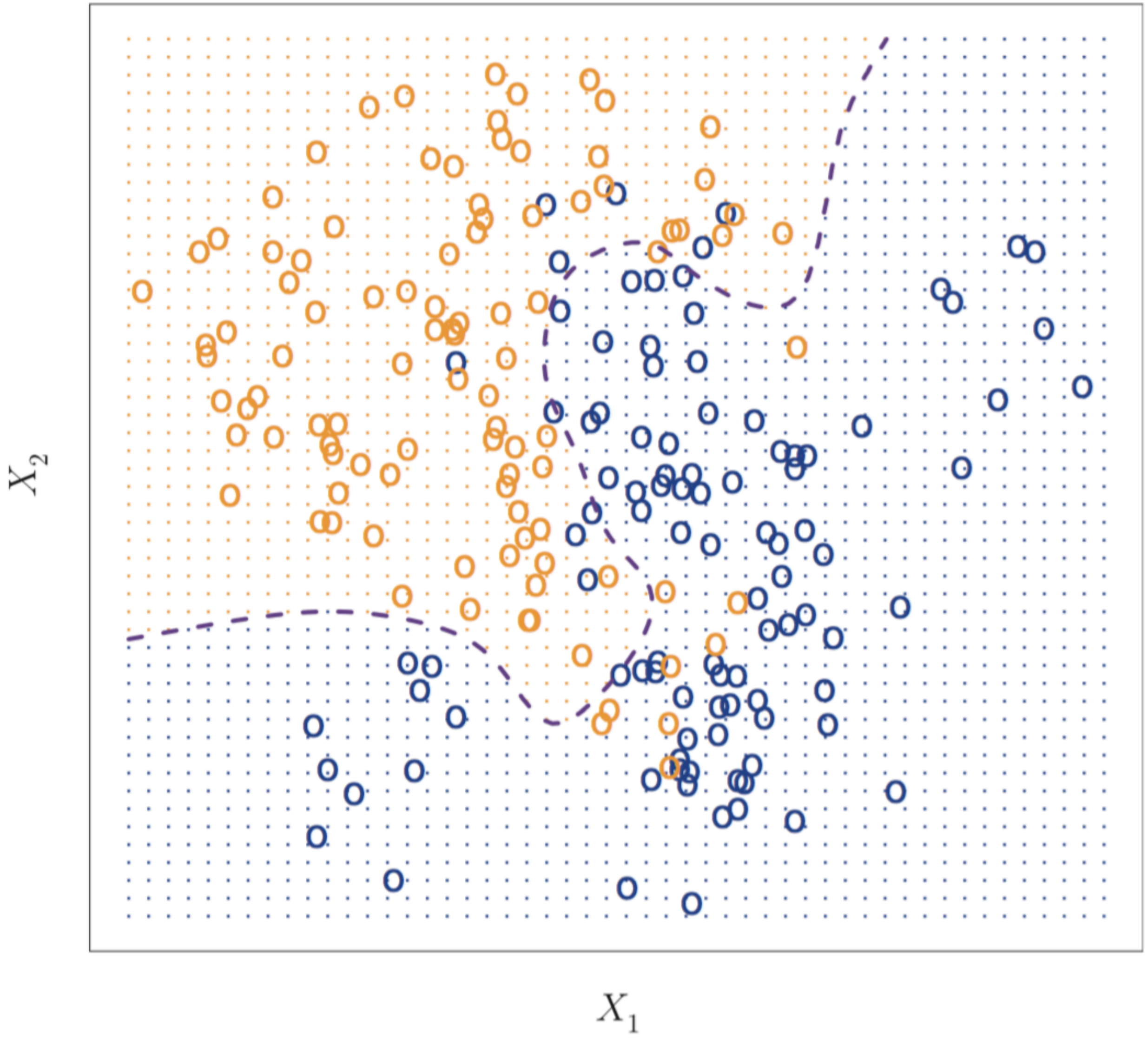
Suppose $P(Y \mid X)$ is known. Then, given an input x_0, we predict the response

$$\hat{y}_0 = \text{argmax}_y \, P(Y = y \mid X = x_0).$$

This Bayes classifier minimizes the expected 0-1 loss:

$$E\left[\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(\hat{y}_i \neq y_i)\right]$$

The minimum expected 0-1 loss (the best we can hope for) is the **Bayes error rate** $1 - E[\text{argmax}_y \, P(Y = y|X)]$. It is the analogon of the irreducible error in regression.



- Suppose there are only two possible response values, say *class 1* or *class 2.*
- The Bayes classifier corresponds to predicting class 1 if

$$\Pr(Y = 1|X = x_0) > 0.5$$

and *class 2* otherwise.
- The *Bayes error rate*

$$1 - E\left(\max_j \Pr(Y = j|X)\right)$$

where the expectation averages the probability over all possible values of X.

K-Nearest Neighborhood

- For real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible.
- Given a positive integer K and a test observation x_0, the KNN classifier

  1. Identifies the K points in the training data closest to x_0, represented by $\mathcal{N}_0$

  2. Estimates

  $$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

  3. Applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

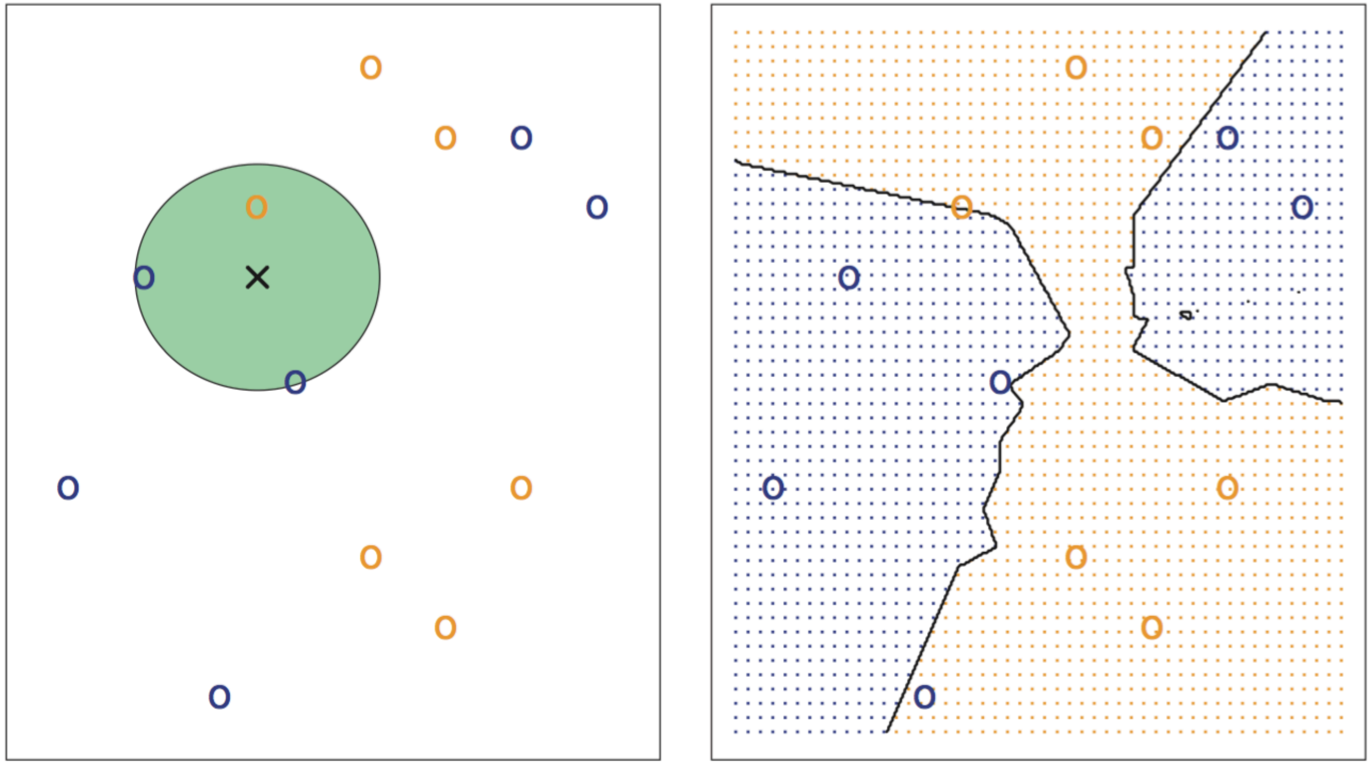


FIGURE 2.14. *The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations.* Left: *a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue.* Right: *The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.*

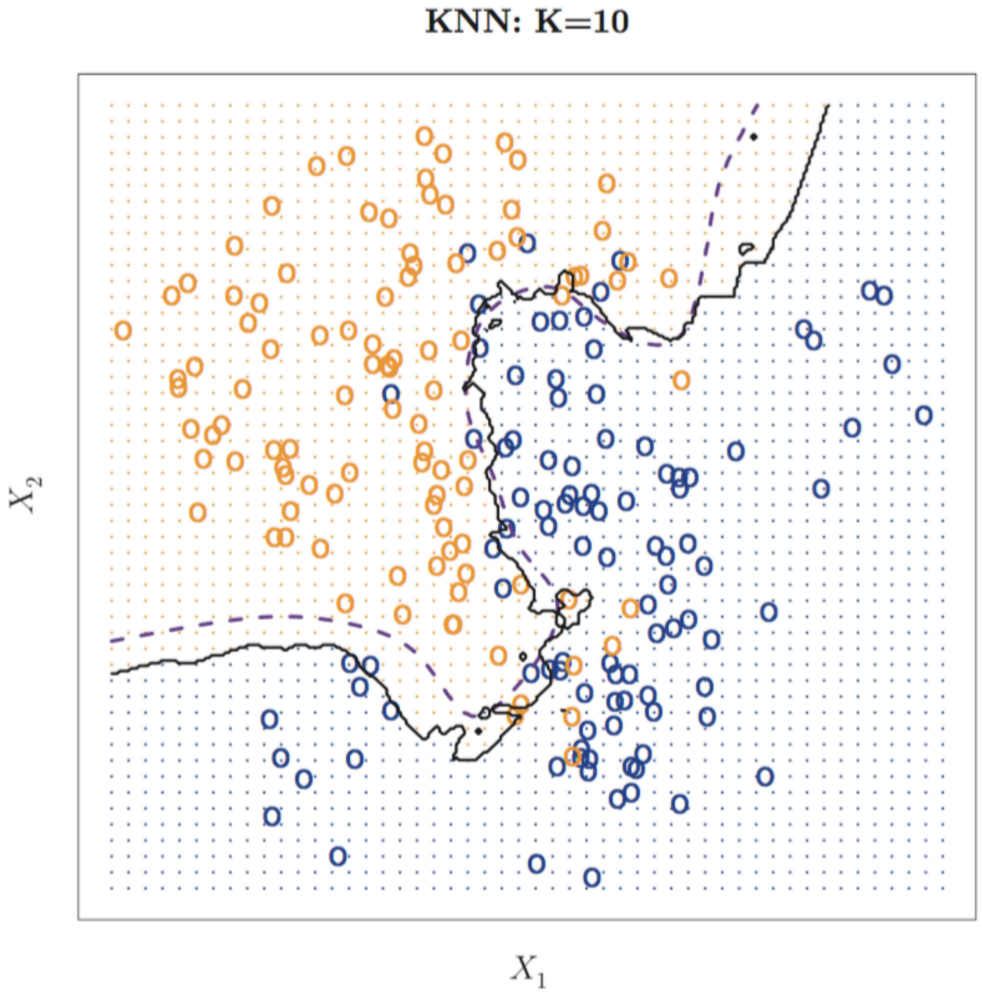


FIGURE 2.15. *The black curve indicates the KNN decision boundary on the data from Figure 2.13, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.*

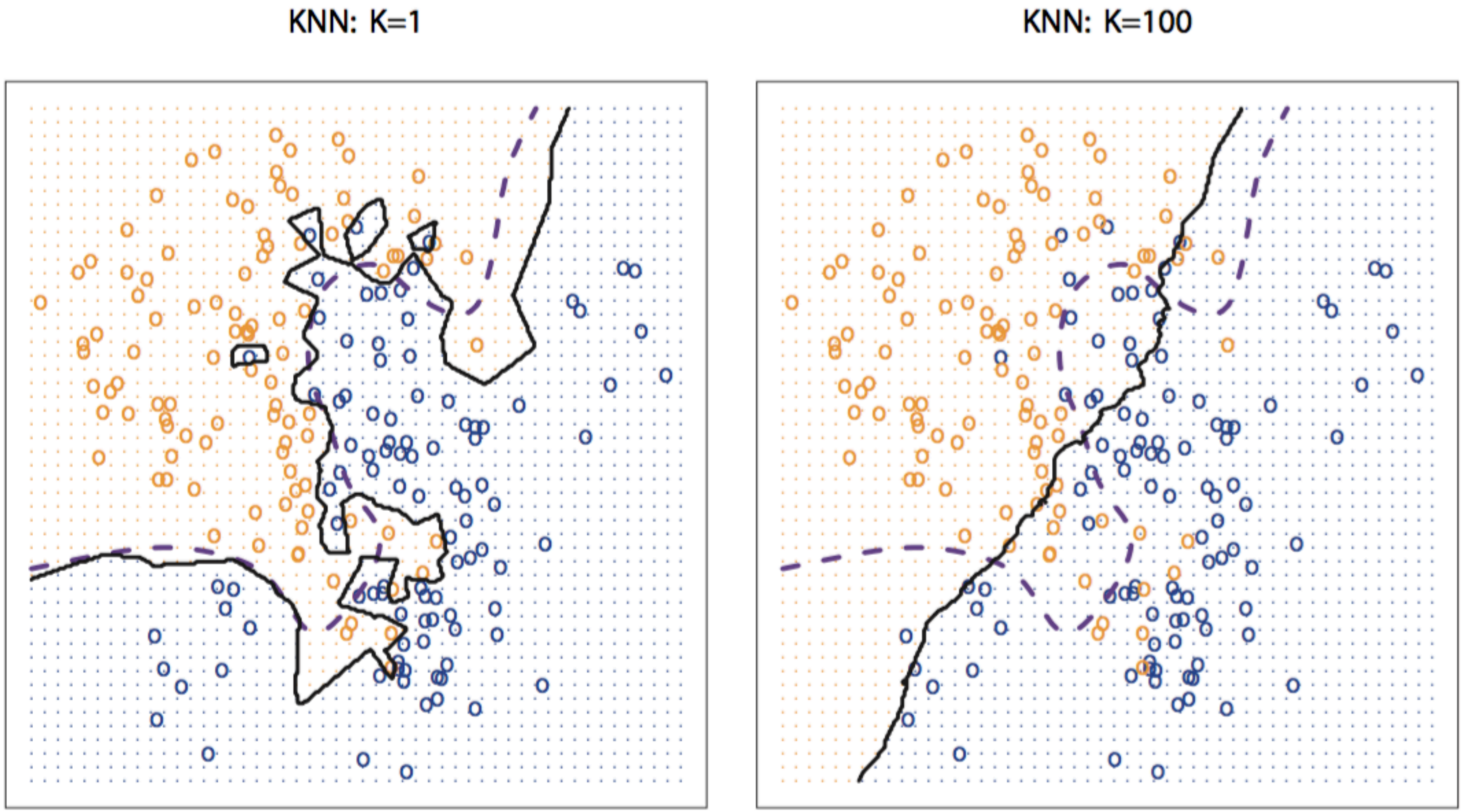
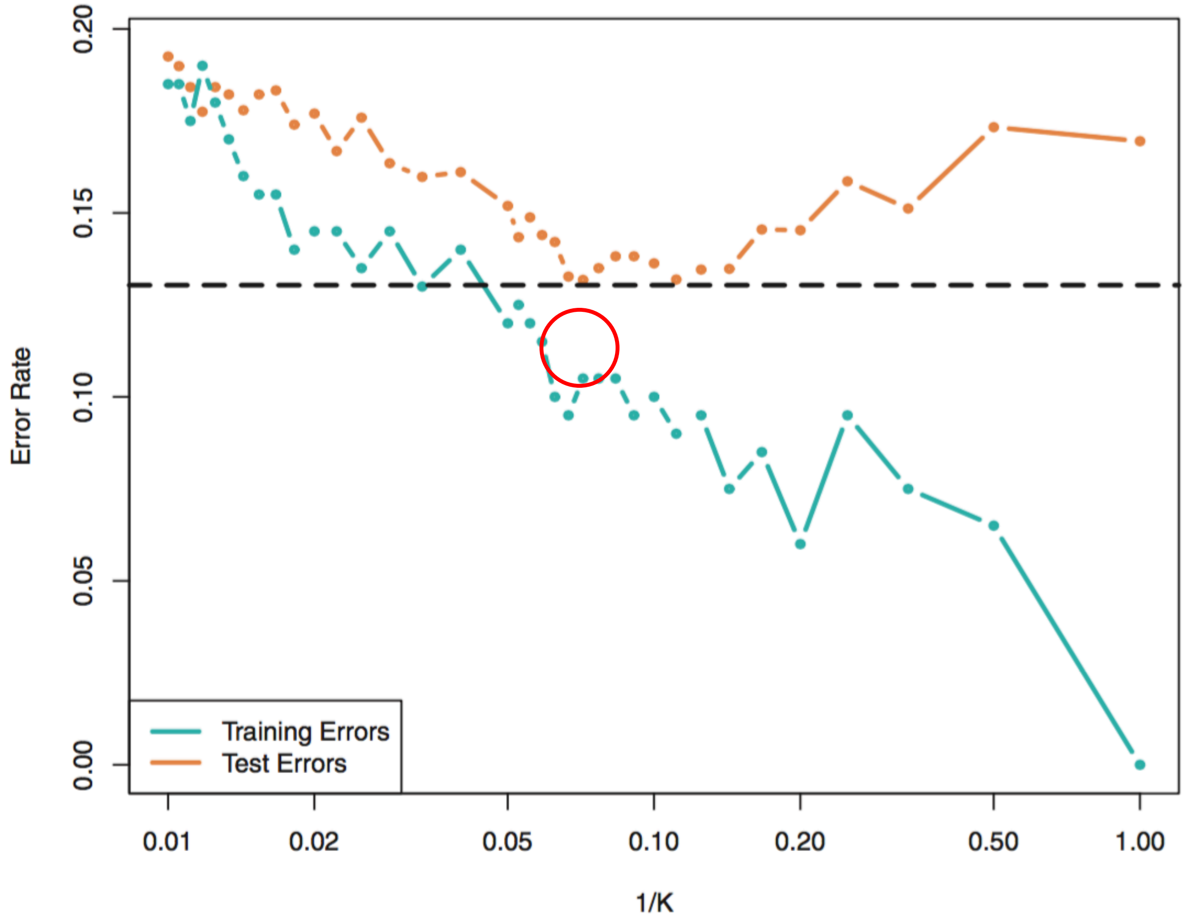


FIGURE 2.16. *A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.*



The black horizontal dotted line implies the Bayes error rate.

- As K grows, it becomes less flexible and a decision boundary gets closer to linear.
- As K grows, it becomes a low-variance but high-bias classifier. Why?
- The bias-variance tradeoff: U-shape in the test error