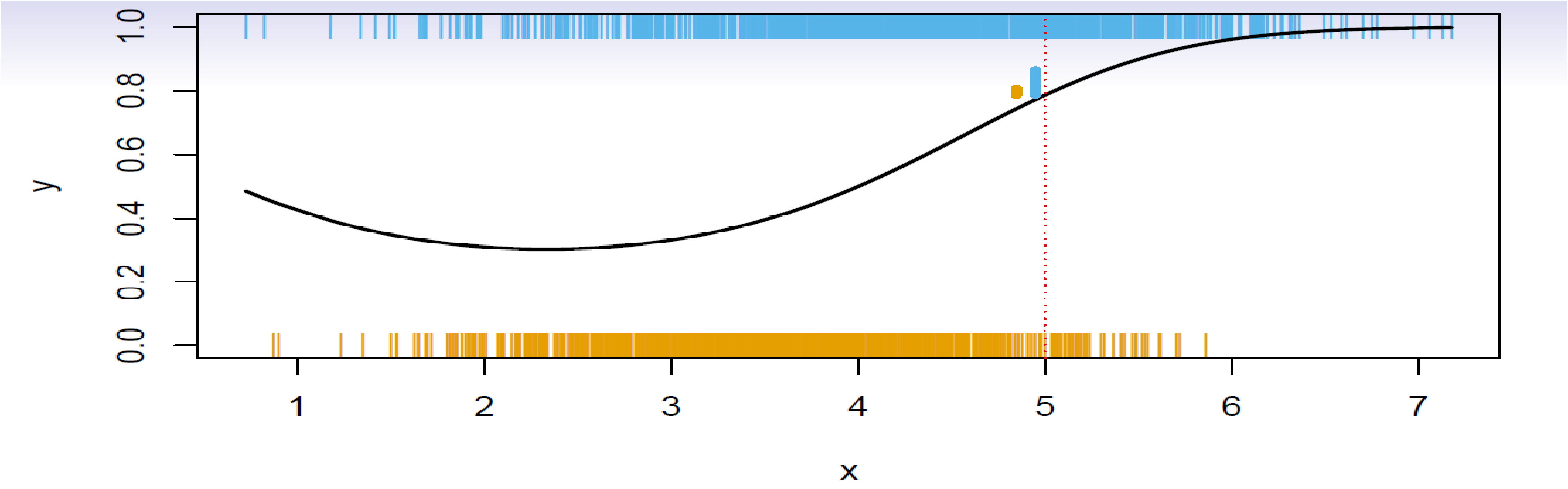
Classification Problems

Here the response variable Y is qualitative — e.g. email is one of C = (spam,ham) (ham=good email), digit class is one of C = {0,1,...,9}. Our goals are to:

- Build a classifier $C(X)$ that assigns a class label from C to a future unlabeled observation X.
- Assess the uncertainty in each classification
- Understand the roles of the different predictors among

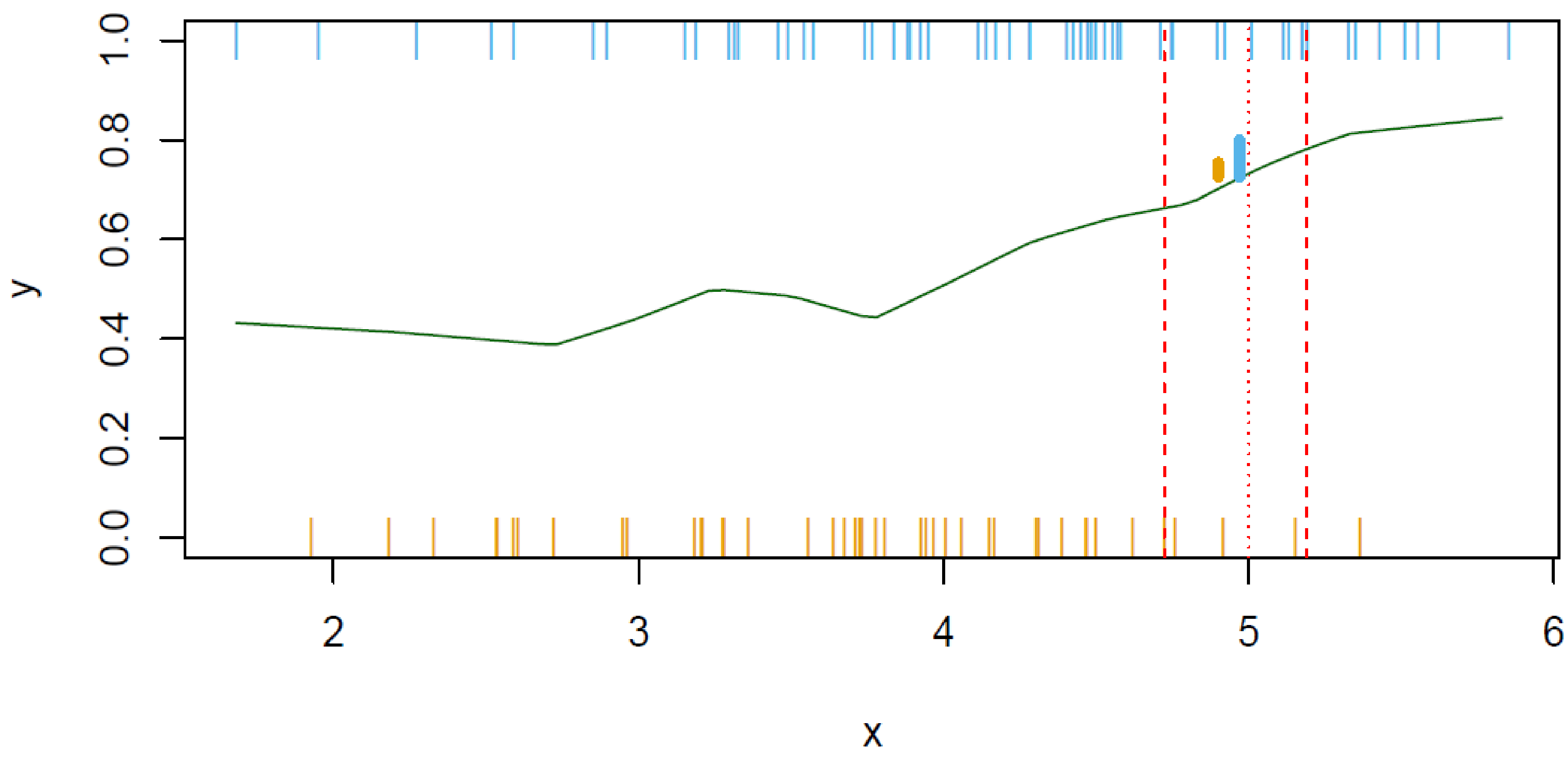
$$X = (X_1, X_2, \dots, X_p).$$



- Is there an ideal $C(X)$? Suppose the K elements in C are numbered $1,2,\ldots,K$. Let
- $p_k(x) = \Pr(Y = k|X = x), k = 1,2,\ldots,K.$

These are the *conditional class probabilities* at x; e.g. see little barplot at x = 5. Then the *Bayes optimal classifier* at x is

$C(x) = j$ if p_k (x) = max$\{p_k(x), p_k(x), \ldots, p_k(x)\}$



- Nearest-neighbor averaging can be used as before. Also breaks down as dimension grows. However, the impact on $\hat{C}(x)$ is less than on $\hat{p}_k(x), k = 1, \ldots, K$.

Classification: some details

- Typically we measure the performance of $\hat{C}(x)$ using the misclassification error rate: $\text{Err}_{\text{Te}} = \text{Ave}_{\text{i}\in\text{Te}}\, I\left[y_i \neq \hat{C}(x_i)\right]$
- The Bayes classifier (using the true $p_k(x)$) has smallest error (in the population).
- Support-vector machines build structured models for $C(x)$.
- We will also build structured models for representing the $p_k(x)$. e.g. Logistic regression, generalized additive models.

The Classification Setting

- For a regression problem, we used the MSE (mean squared error) to assess the accuracy of the statistical learning method
- For a classification problem we can use the error rate i.e.

$$Error\ Rate = \sum_{i=1}^{n} I(y_i \neq \hat{y}_i) / n$$

- $I(y_i \neq \hat{y}_i)$ is an indicator function, which will give 1 if the condition $(y_i \neq \hat{y}_i)$ is correct, otherwise it gives a 0.
- Thus the error rate represents the fraction of incorrect classifications, or misclassifications

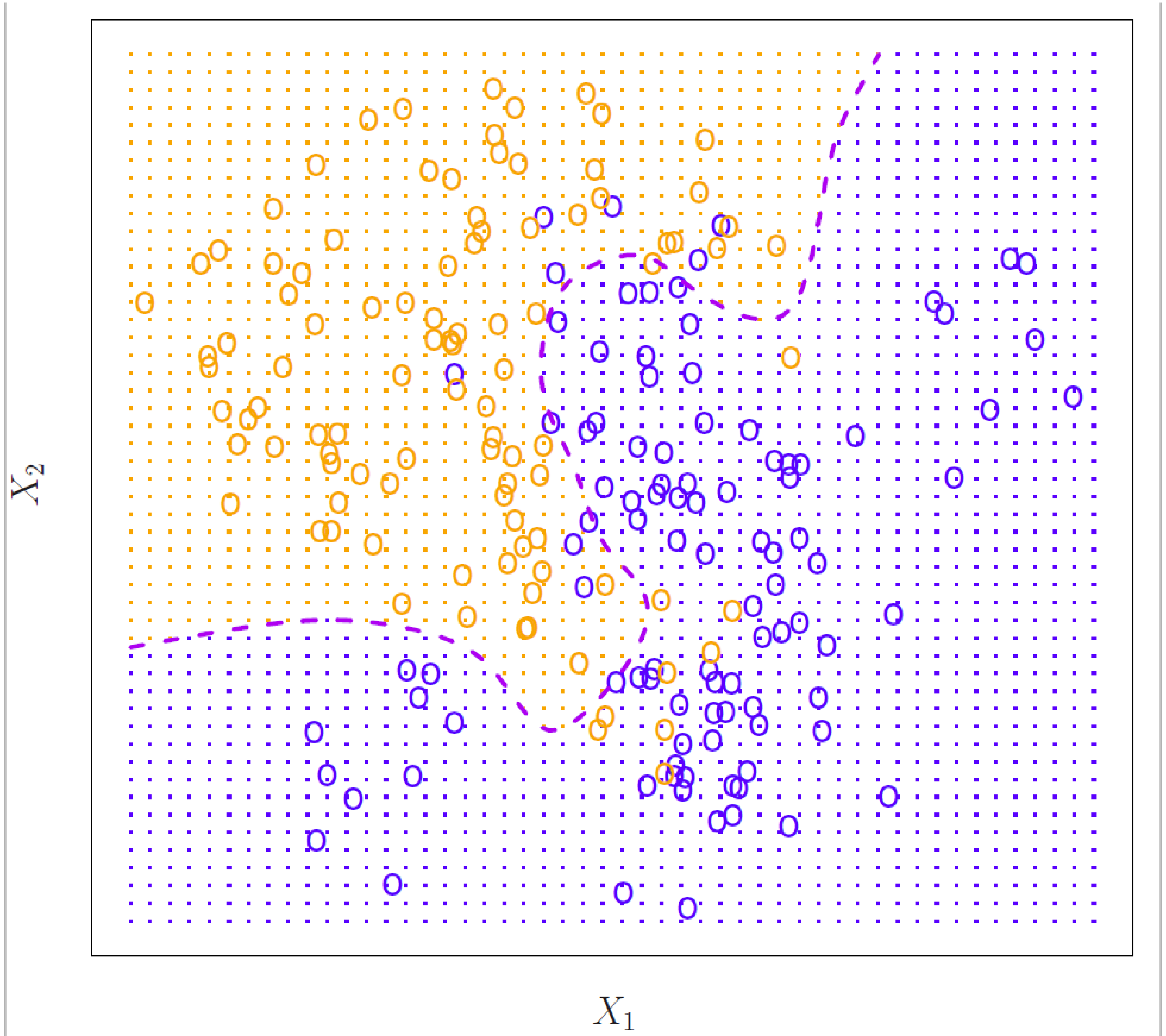
Bayes Error Rate

- The Bayes error rate refers to the lowest possible error rate that could be achieved if somehow we knew exactly what the "true" probability distribution of the data looked like.
- On test data, no classifier (or stat. learning method) can get lower error rates than the Bayes error rate.
- Of course in real life problems the Bayes error rate can't be calculated exactly.

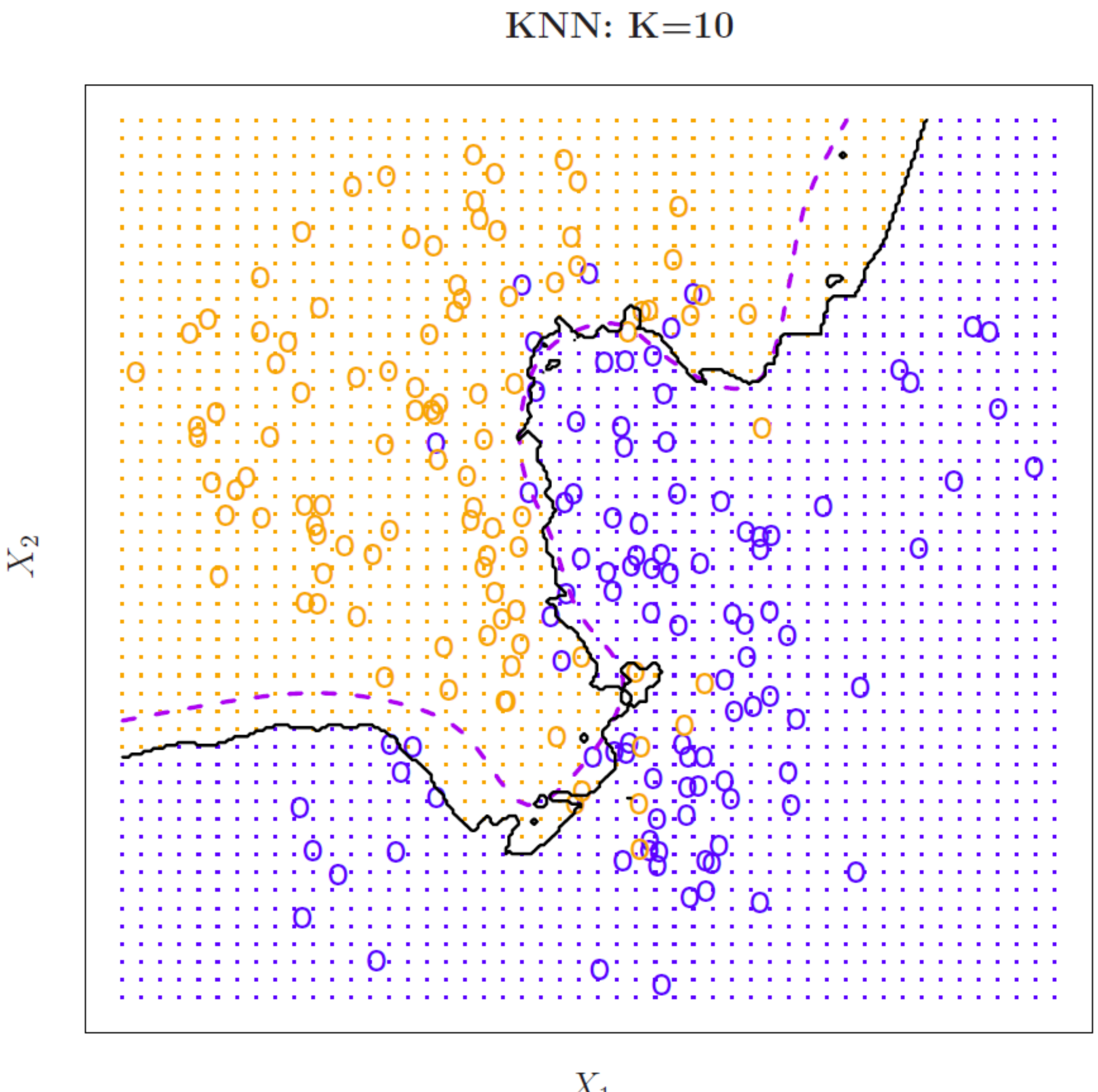
K-Nearest Neighbors (KNN)

- K Nearest Neighbors is a flexible approach to estimate the Bayes Classifier.
- For any given X we find the k closest neighbors to X in the training data, and examine their corresponding Y.
- If the majority of the Y's are orange we predict orange otherwise guess blue.
- The smaller that k is, the more flexible the method will be.

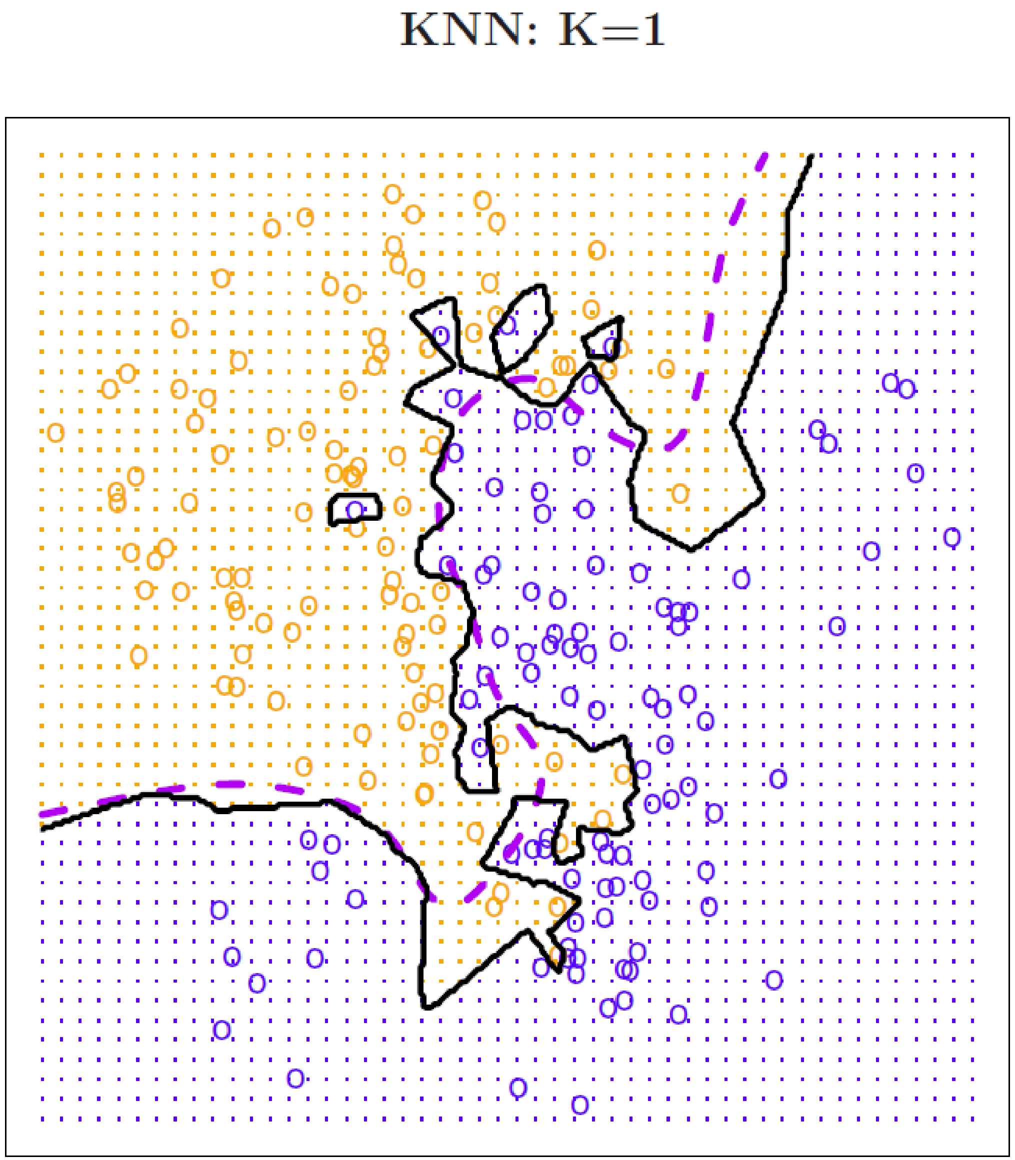
Example: K-nearest neighbors in two dimensions

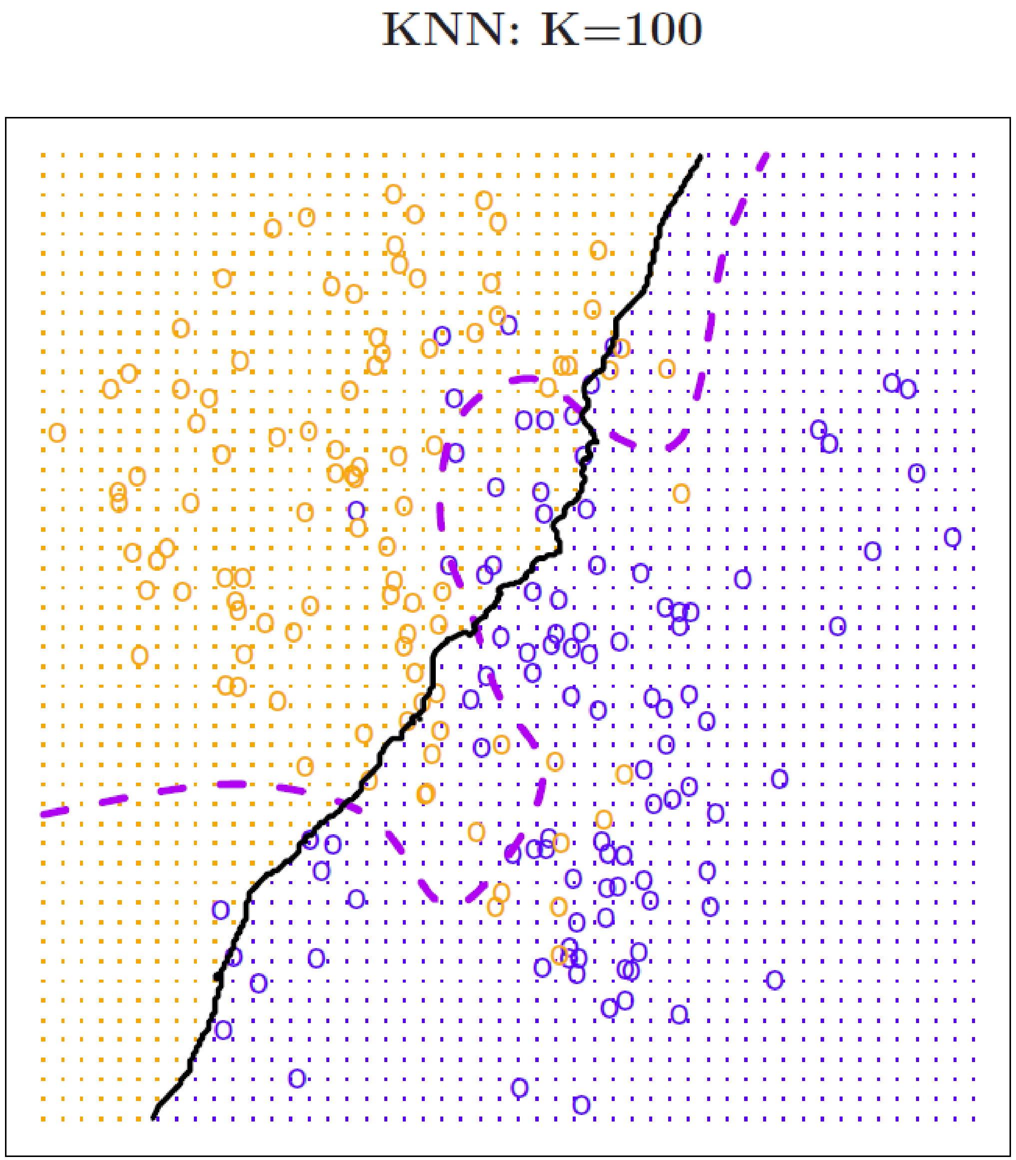


Simulated Data: K = 10



K=1 and K = 100

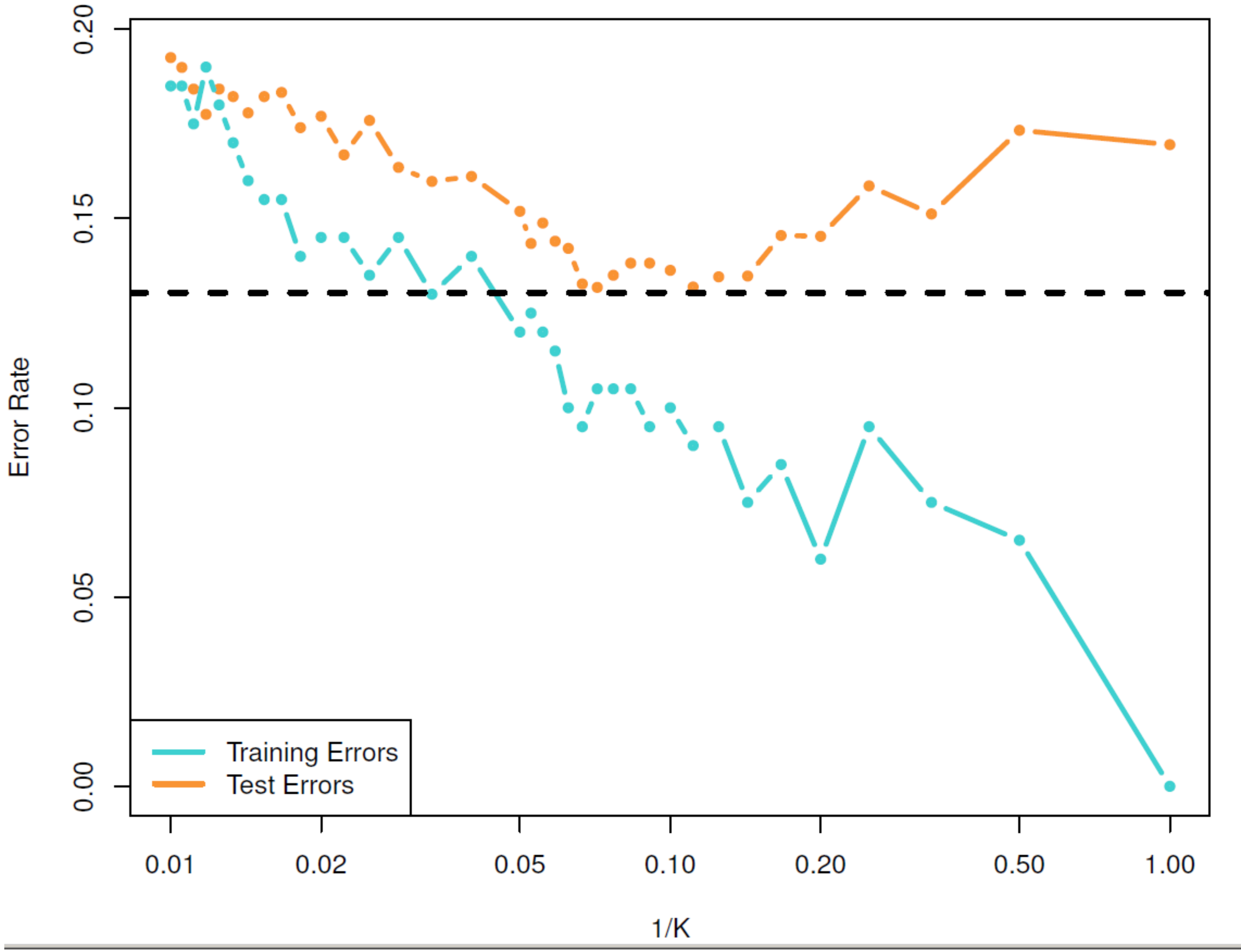




Training vs. Test Error Rates on the Simulated Data

- Notice that training error rates keep going down as k decreases or equivalently as the flexibility increases.

- However, the test error rate at first decreases but then starts to increase again.



A Fundamental Picture

- In general training errors will always decline.
- However, test errors will decline at first (as reductions in bias dominate) but will then start to increase again (as increases in variance dominate)
- We must always keep this picture in mind when choosing a learning method. More flexible/complicated is not always better!