Class 6 – K Nearest Neighbors (KNN)

Classification

Pedram Jahangiry

Fall 2019

Jon M.
Huntsman
School of Business
UtahStateUniversity

Confusion Matrix

		Predictions	
		0 negative	1 positive
Actual	0 negative	TN	FP*
	1 positive	FN**	TP

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

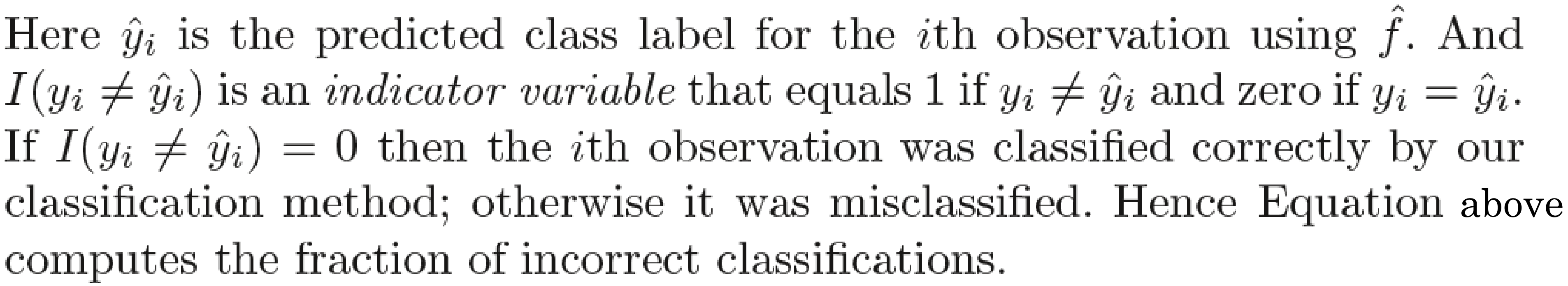
FP*	Type I error
FN**	Type II eror

The Classification Setting

- The most common approach for quantifying the accuracy our estimate f̂ is the training error rate:

The proportion of mistakes that are made if we apply our estimate f̂ to the training observations

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



Here $\hat{y}_i$ is the predicted class label for the ith observation using $\hat{f}$. And $I(y_i \neq \hat{y}_i)$ is an *indicator variable* that equals 1 if $y_i \neq \hat{y}_i$ and zero if $y_i = \hat{y}_i$. If $I(y_i \neq \hat{y}_i) = 0$ then the ith observation was classified correctly by our classification method; otherwise it was misclassified. Hence Equation above computes the fraction of incorrect classifications.

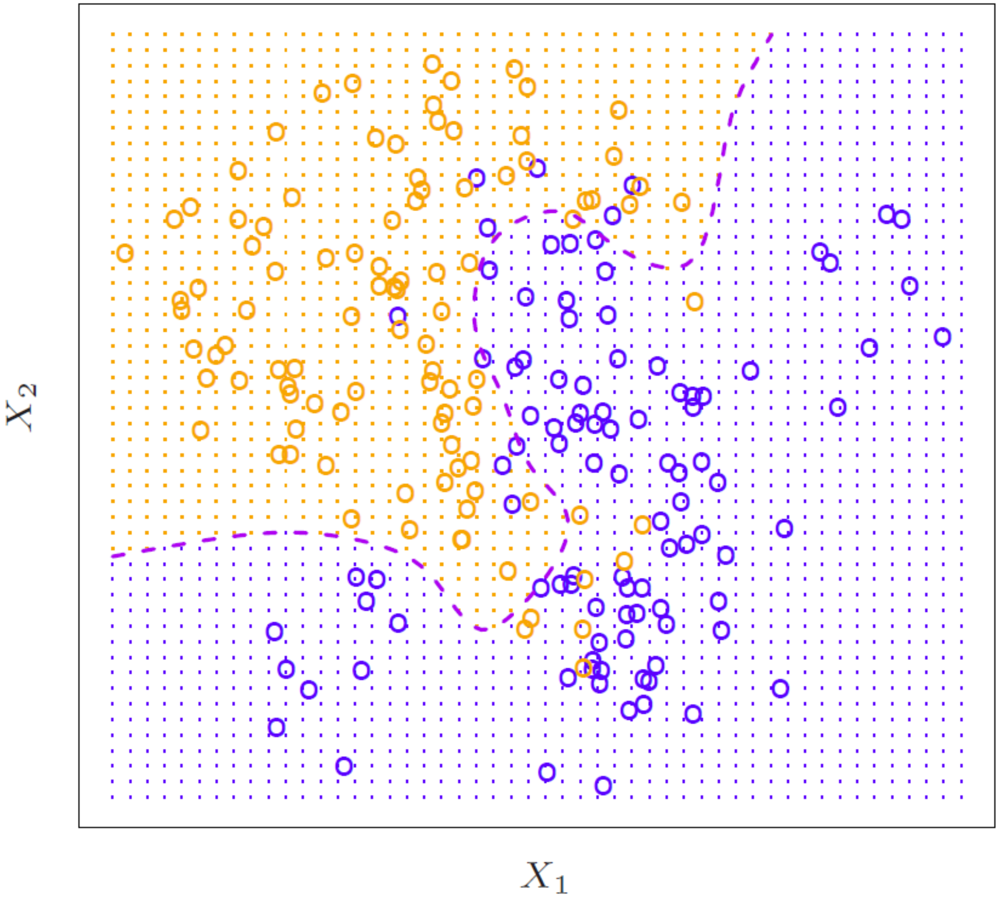
The Classification Setting

- As in the error regression setting, we are most interested in the error rates that result from applying our classifier to test observations that were not used in training.
- A good classifier is one for which the test error is smallest.
- It is possible to show that the test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values.
- In other words, we should simply assign a test observation with predictor vector x_0 to the class j for which, the conditional probability $\mathrm{P}(\mathrm{y} = \mathrm{j} | \mathrm{X} = \mathrm{x}_0)$ is largest.
- This very simple classifier is called the Bayes classifier.
- In a two-class problem the Bayes classifier corresponds to predicting class one if $\mathrm{P}(\mathrm{y} = 1 | \mathrm{X} = \mathrm{x}_0) > 0.5$, and class two otherwise.

Bayes classifier in two dimensions

A simulated data set consisting of 100 *observations in each of two groups, indicated in blue and in orange.*

- *The purple dashed line represents the Bayes decision boundary.*
- *The orange background grid indicates the region in which a test observation will be assigned to the orange class, and*
- *the blue background grid indicates the region in which a test observation will be assigned to the blue class.*



K-Nearest Neighbors

- In theory we would always like to predict qualitative responses using the Bayes classifier

- In practice we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible.

- Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods.

- Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability. One such method is the K-nearest neighbors (KNN) classifier.

K-Nearest Neighbors

Given a positive integer K and a test observation x_0, the KNN classifier:

- First identifies the K points in the training data that are closest to x_0, represented by N_0.
- Then estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j:

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

- Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

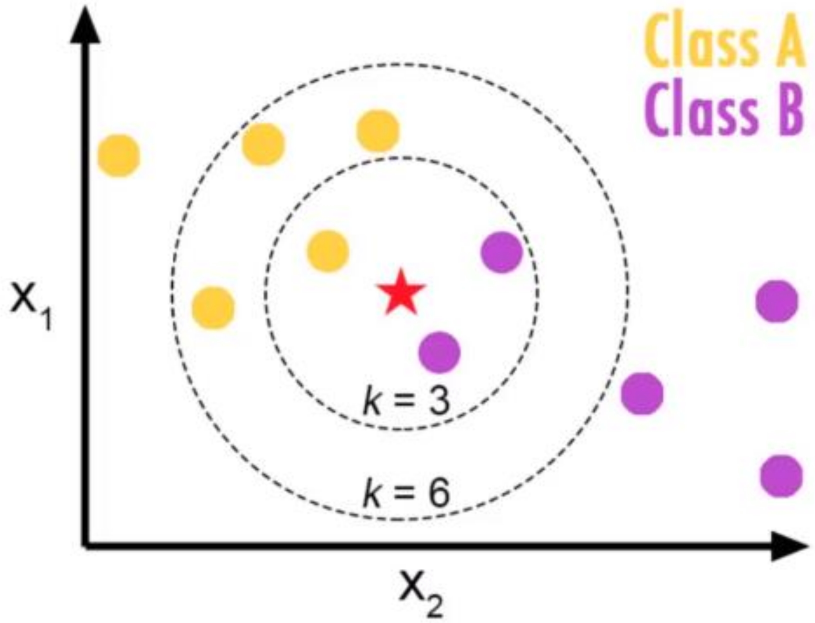
K-Nearest Neighbors

Pros

- *Simple*
- *Works with any number of classes*
- *Few parameters (K, distance metric)*

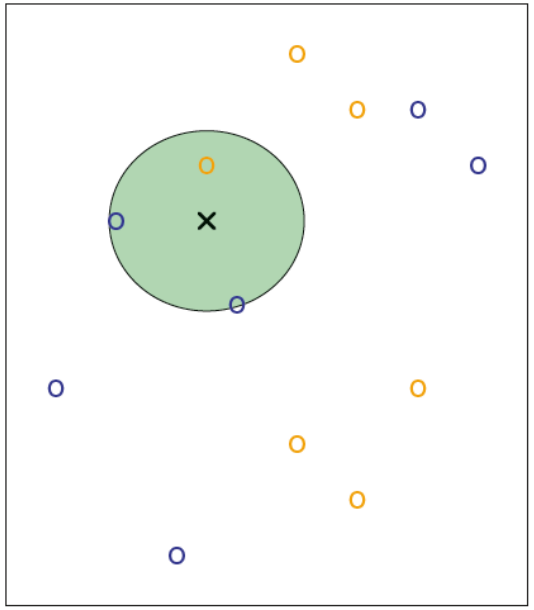
Cons

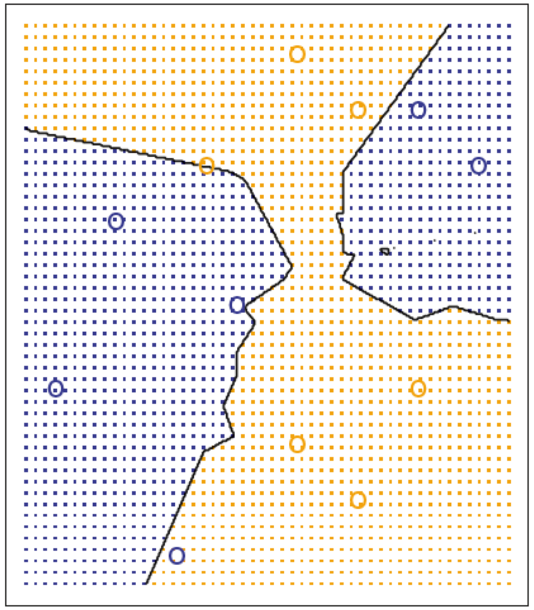
- *Curse of dimensionality*
- *Not good for categorical features*



K-Nearest Neighbors

The KNN approach, using K = 3, is illustrated in a simple situation with six blue observations and six orange observations.





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.

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.

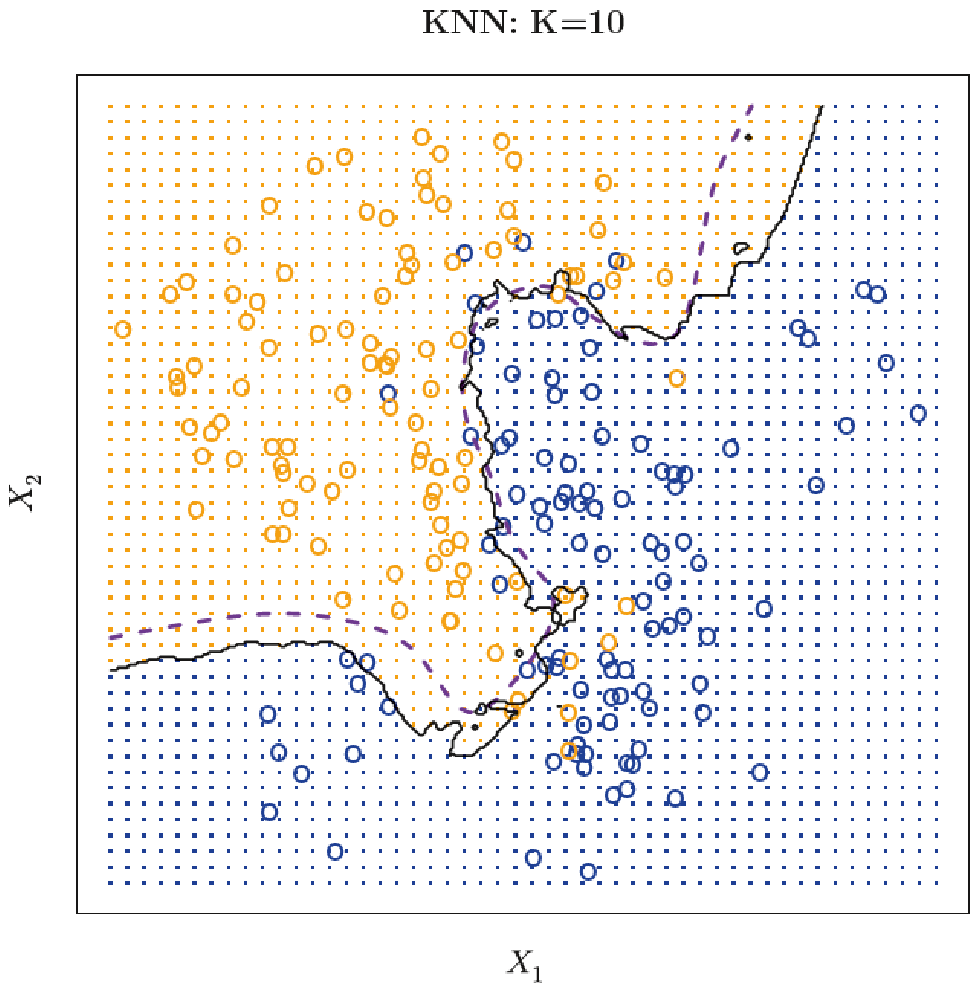
K-Nearest Neighbors

- *The black curve indicates the KNN decision boundary on the data using K = 10.*

- *The Bayes decision boundary is shown as a purple dashed line.*

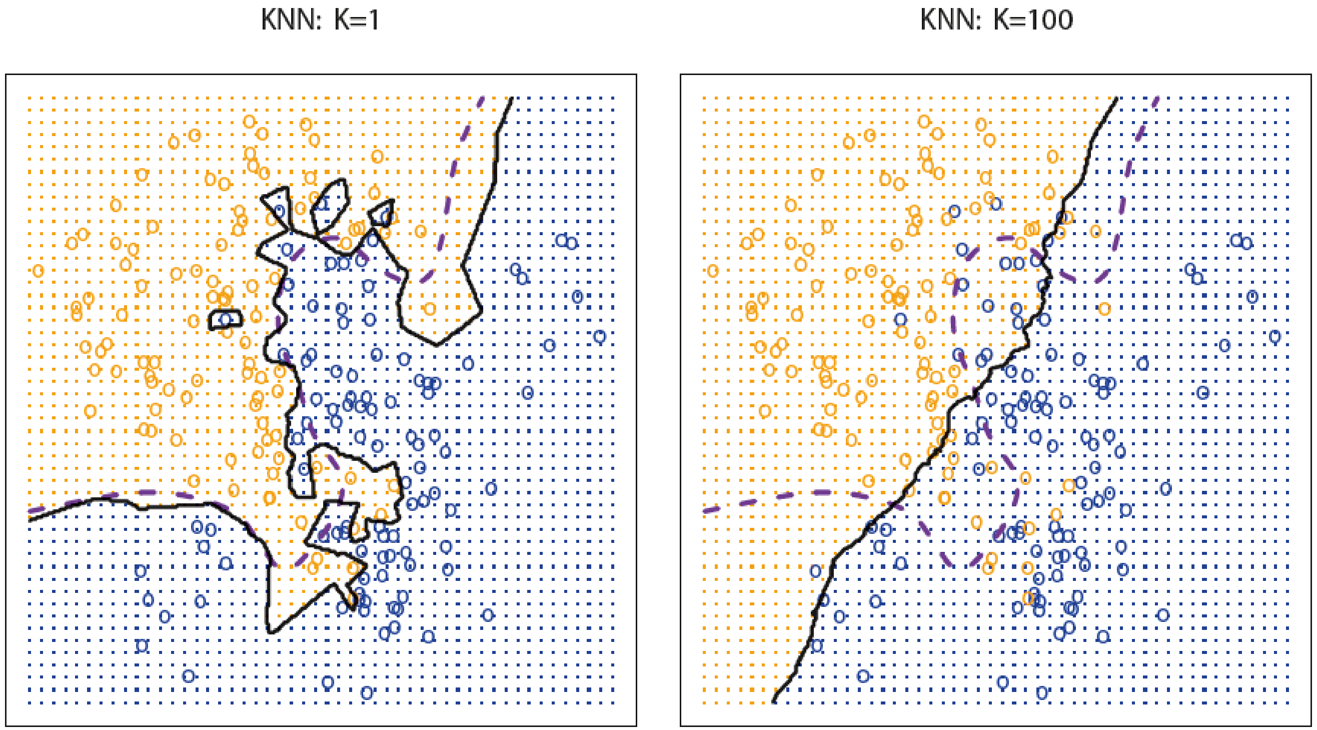
- *Despite the fact that KNN is a very simple approach, it can often produce classifiers that are surprisingly close to the optimal Bayes classifier.*

- *The test error rate using KNN is 0.1363, which is close to the Bayes error rate of 0.1304.*



Choice of K

- The choice of *K* has a drastic effect on the KNN classifier obtained.

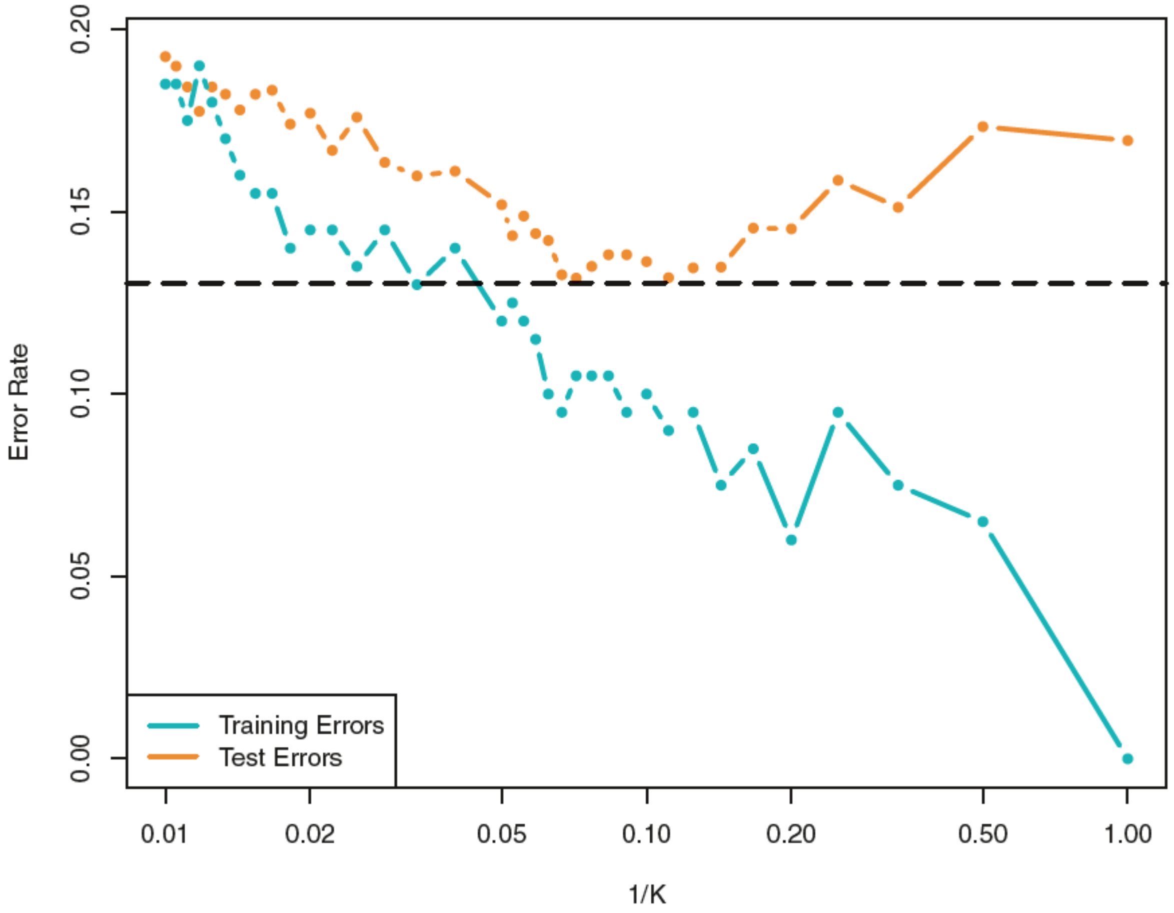


KNN: K=1

KNN: K=100

- What can you say about the bias variance trade off and the flexibility of the model?
- What about the linearity / non-linearity of the classifier (KNN)?

Choice of K



- The KNN training / Test error rate shown as the level of flexibility (assessed using 1/K) increases, or equivalently as the number of neighbors K decreases.

- The black dashed line indicates the Bayes error rate.

- The jumpiness of the curves is due to the small size of the training data set.

Optimal level of K by using cross validation

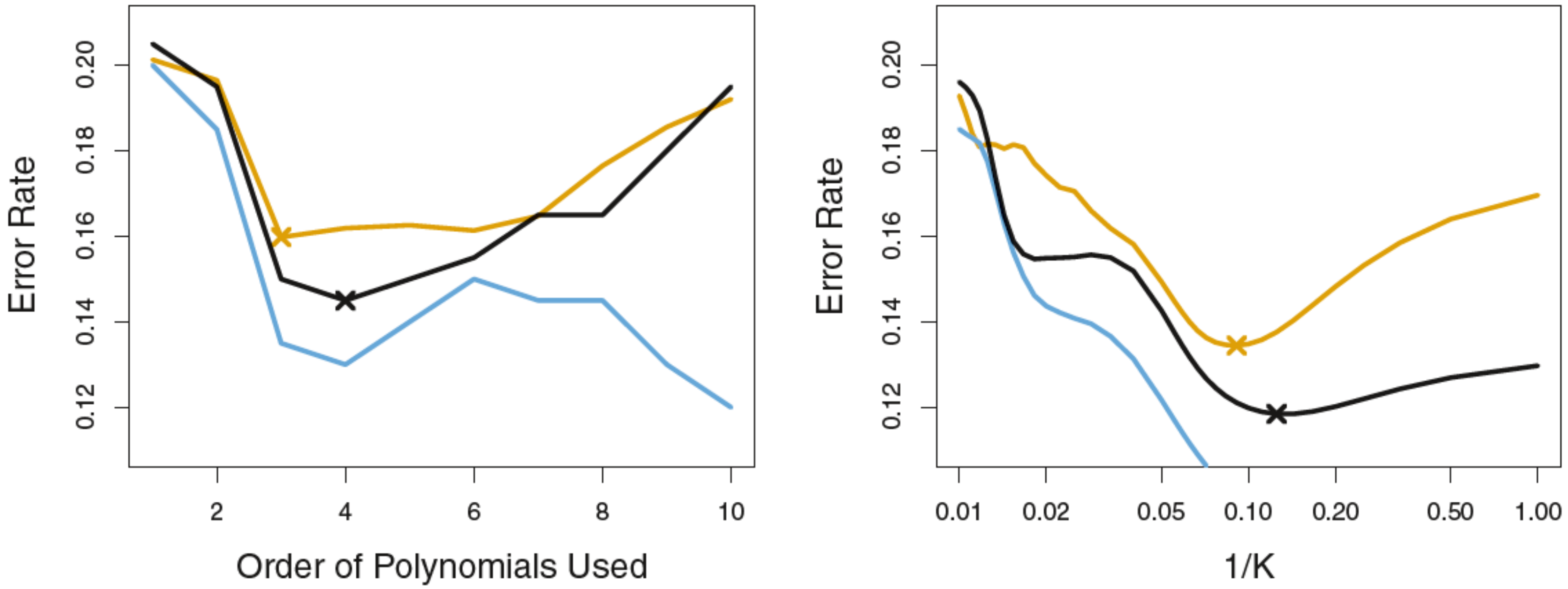


FIGURE 5.8. *Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7.* Left: *Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis.* Right: *The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.*

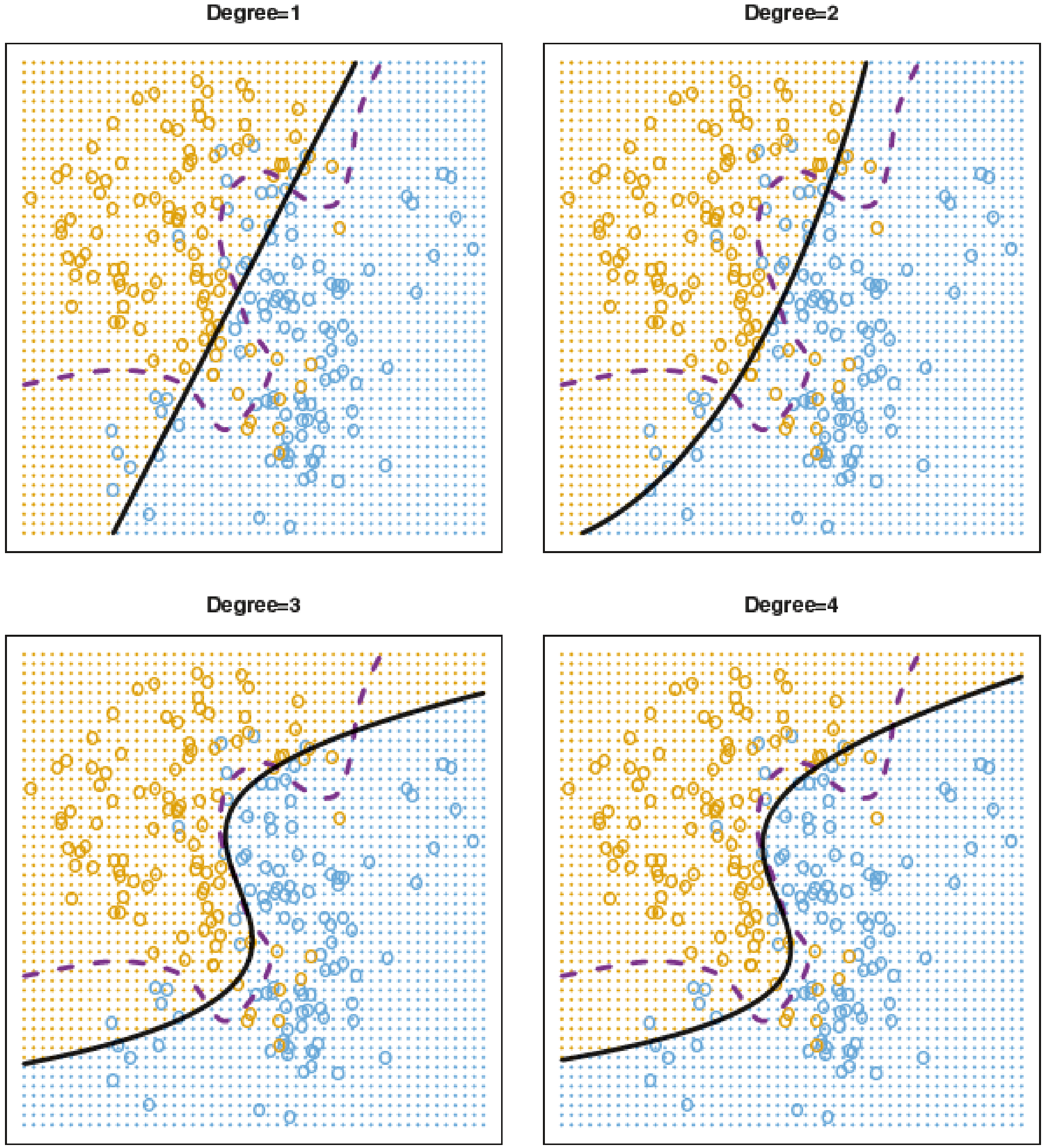


FIGURE 5.7. *Logistic regression fits on the two-dimensional classification data displayed in Figure 2.13. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1–4) logistic regressions are displayed in black. The test error rates for the four logistic regression fits are respectively* 0.201, 0.197, 0.160, *and* 0.162, *while the Bayes error rate is* 0.133.