Now the knn() function can be used to predict the default variable for the test data.

```
> set.seed(1)
> knn_pred <- knn(train_X, test_X, train_default, k = 1)
> table(knn_pred, test_default)
        test_default
knn_pred  No Yes
     No  944  25
     Yes  20  11
> mean(knn_pred == test_default)
[1] 0.955
```

Below, we repeat the analysis using $K = 12$.

```
> set.seed(1)
> knn_pred <- knn(train_X, test_X, train_default, k = 12)
> table(knn_pred, test_default)
        test_default
knn_pred  No Yes
     No  961  26
     Yes   3  10
> mean(knn_pred == test_default)
[1] 0.971
```

The results have improved slightly. But increasing K further turns out to provide no further improvements.

An Analytical Comparison of Classification Methods

We now perform an analytical (or mathematical) comparison of LDA, QDA, naive Bayes, and logistic regression.

- We consider these approaches in a setting with K classes, so that we assign an observation to the class that maximizes $\Pr(Y = k|X = x)$.

- Equivalently, we can set K as the baseline class and assign an observation to the class that maximizes

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right)$$

for $k=1,...,K$.

- Examining the specific form of formula above for each method provides a clear understanding of their similarities and differences.

- First, for LDA, we can make use of Bayes' Theorem as well as the assumption that the predictors within each class are drawn from a multivariate normal density with class-specific mean and shared covariance matrix in order to show that

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j,$$

where $a_k = \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K)$ and b_{kj} is the jth component of $\mathbf{\Sigma}^{-1}(\mu_k-\mu_K)$.

- Hence LDA, like logistic regression, assumes that the log odds of the posterior probabilities is linear in x.

- Using similar calculations, in the QDA setting

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{kjl}x_jx_l,$$

where a_k, b_{kj}, and c_{kjl} are functions of π_k, π_K, μ_k, μ_K, $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_K$.

- Again, as the name suggests, QDA assumes that the log odds of the posterior probabilities is quadratic in x.

- In the naive Bayes setting, recall that, $f_k(x)$ is modeled as a product of p one-dimensional functions $f_{kj}(x_j)$ for $j=1,...,p$. Hence,

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = a_k + \sum_{j=1}^{p} g_{kj}(x_j),$$

where $a_k = \log\left(\frac{\pi_k}{\pi_K}\right)$ and $g_{kj}(x_j) = \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right)$.

Inspection of formulas above yields the following observations about LDA, QDA, and naive Bayes:

i. LDA is a special case of QDA with $c_{kjl} = 0$ for all $j = 1, ..., p$, $l = 1, ..., p$, and $k = 1, ..., K$.

ii. Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kj}(x_j) = b_{kj}x_j$. In particular, this means that LDA is a special case of naive Bayes!

iii. If we model $f_{kj}(x_j)$ in the naive Bayes classifier using a one-dimensional Gaussian distribution $N(\mu_{kj}, \sigma_j^2)$, then we end up with $g_{kj}(x_j) = b_{kj}x_j$ where $b_{kj} = (\mu_{kj} - \mu_{Kj})/\sigma_j^2$. In this case, naive Bayes is actually a special case of LDA with $\mathbf{\Sigma}$ restricted to be a diagonal matrix with jth diagonal element equal to σ_j^2.

iv. Neither QDA nor naive Bayes is a special case of the other.

Note: None of these methods uniformly dominates the others: in any setting, the choice of method will depend on the true distribution of the predictors in each of the K classes, as well as other considerations, such as the values of n and p.

How does logistic regression tie into this story?

- Recall that multinomial logistic regression takes the form

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \sum_{j=1}^{p} \beta_{kj}x_j.$$

- This is identical to the linear form of LDA: in both cases,

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right)$$

is a linear function of the predictors.

- In LDA, the coefficients in this linear function are functions of estimates for π_k, π_K, μ_k, μ_K, and $\mathbf{\Sigma}$ obtained by assuming that $X_1, ..., X_p$ follow a normal distribution within each class.
- By contrast, in logistic regression, the coefficients are chosen to maximize a likelihood function.
- Thus, we expect LDA to outperform logistic regression when the normality assumption (approximately) holds, and we expect logistic regression to perform better when it does not.

An Empirical Comparison of Classification Methods

We now compare the empirical (practical) performance of logistic regression, LDA, QDA, naive Bayes, and KNN.

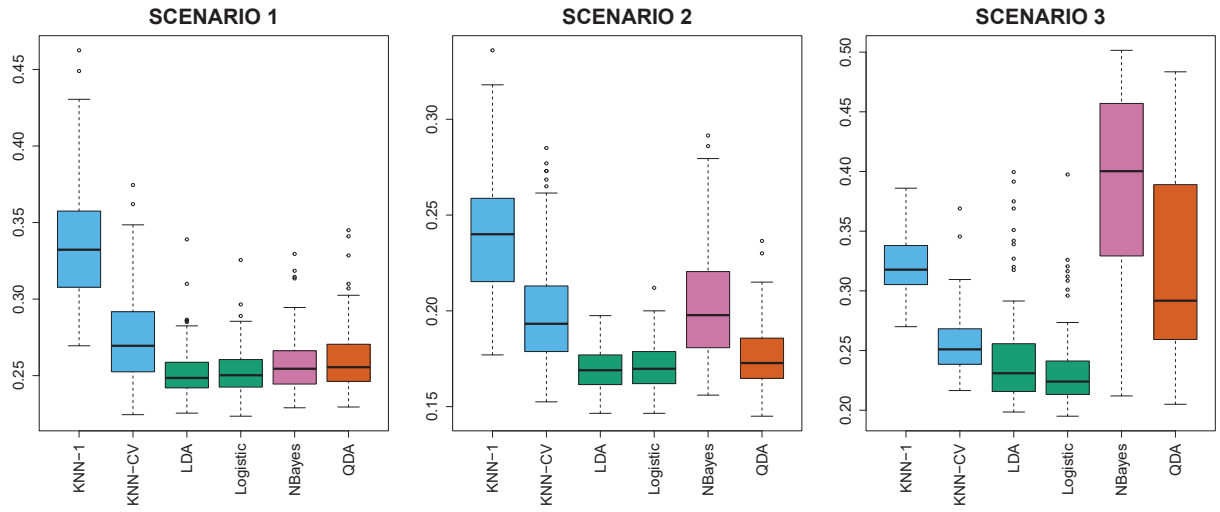
We generated data from six different scenarios, each of which involves a binary (two-class) classification problem. In three of the scenarios, the Bayes decision boundary is linear, and in the remaining scenarios it is non-linear.

In each of the six scenarios, there were $p = 2$ quantitative predictors. The scenarios were as follows:

Scenario 1: There were 20 training observations in each of two classes. The observations within each class were uncorrelated random normal variables with a different mean in each class.

Scenario 2: Details are as in Scenario 1, except that within each class, the two predictors had a correlation of -0.5.

Scenario 3: As in the previous scenario, there is substantial negative correlation between the predictors within each class. However, this time we generated X_1 and X_2 from the t-distribution, with 50 observations per class.

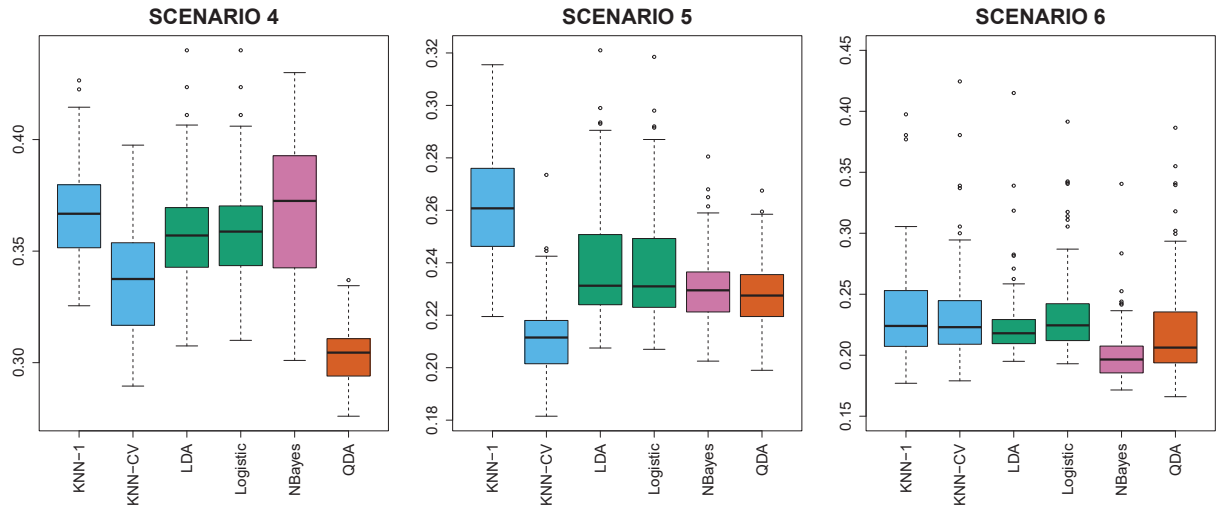


Boxplots of the test error rates for each of the linear scenarios.

Scenario 4: The data were generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and correlation of -0.5 between the predictors in the second class.

Scenario 5: The data were generated from a normal distribution with uncorrelated predictors. Then the responses were sampled from the logistic function applied to a complicated non-linear function of the predictors.

Scenario 6: The observations were generated from a normal distribution with a different diagonal covariance matrix for each class. However, the sample size was very small: just $n = 6$ in each class.



Boxplots of the test error rates for each of the non-linear scenarios.

These examples show that no one method will dominate the others in every situation.

Inspection of these six examples yields the following observations about the classifications methods:

i. When the true decision boundaries are linear, then the LDA and logistic regression approaches will tend to perform well.

ii. When the boundaries are moderately non-linear, QDA or naive Bayes may give better results.

iii. Finally, for much more complicated decision boundaries, a non-parametric approach such as KNN can be superior. But the level of smoothness for a non-parametric approach must be chosen carefully.