4.5 A Comparison of Classification Methods

Zongyi Liu

2023-05-20

4.5 A Comparison of Classification Methods

In this chapter we have considered three classification approaches: logistic, LDA, and QDA.

In the LDA framework, the log odds is given by

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right)=\log\left(\frac{p_1(x)}{p_2(x)}\right)=c_0+c_1x,$$

Where c_0 and c_1 are functions of μ_1, μ_2, and σ^2. From previous inductions, we know that in logistic regression,

$$\log\left(\frac{p_1}{1-p_1}\right)=\beta_0+\beta_1x.$$

Both equations above are linear functions of x. Hence, both logistic re- gression and LDA produce linear decision boundaries. The only difference between the two approaches lies in the fact that β_0 and β_1 are estimated using maximum likelihood, whereas c_0 and c_1 are computed using the estimated mean and variance from a normal distribution.

LDA assumes that the observations are drawn from a Gaussian distribution with a common covariance matrix in each class, and so can provide some improvements over logistic regression when this assumption approximately holds. Conversely, logistic regression can outperform LDA if these Gaussian assumptions are not met.

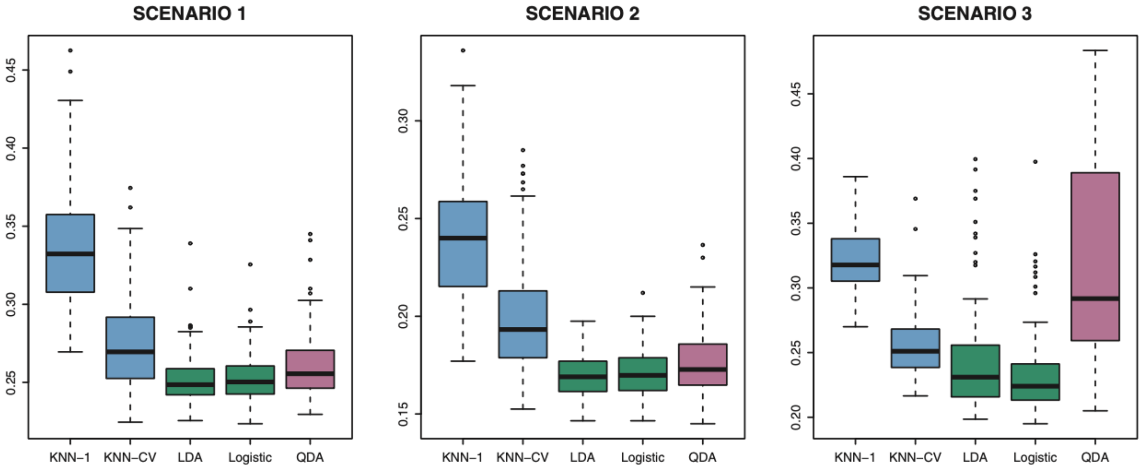
KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary. There- fore, we can expect this approach to dominate LDA and logistic regression when the decision boundary is highly non-linear.

We can get six scnarioes to compare those methods

- **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. The left-hand panel of Figure 4.10 shows that LDA performed well in this setting, as one would expect since this is the model assumed by LDA. KNN performed poorly because it paid a price in terms of variance that was not offset by a reduction in bias. QDA also performed worse than LDA, since it fit a more flexible classifier than necessary. Since logistic regression assumes a linear decision boundary, its results were only slightly inferior to those of LDA.

- **Scenario 2**: Details are as in Scenario 1, except that within each class, the two predictors had a correlation of -0.5. The center panel of Figure 4.10 indicates little change in the relative performances of the methods as compared to the previous scenario.

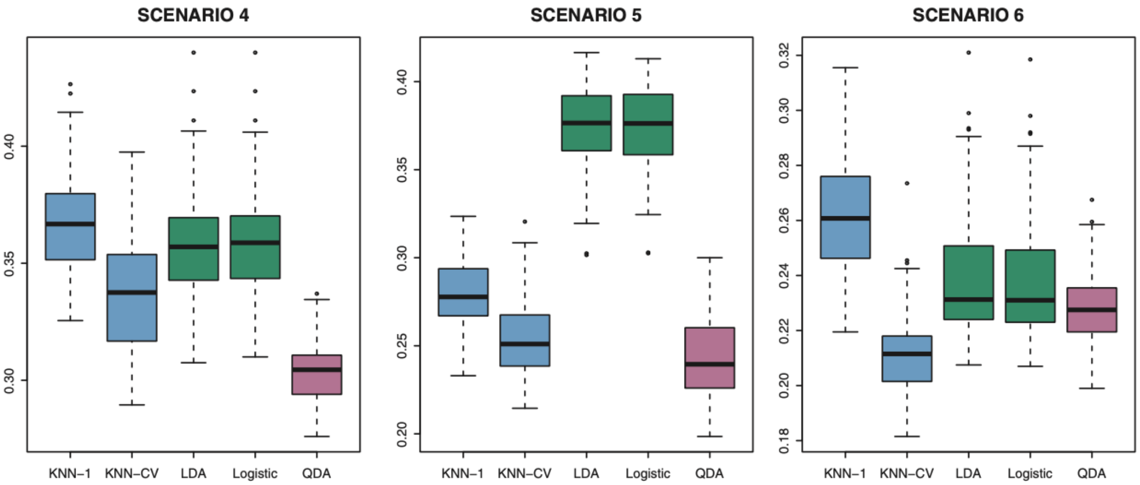
- **Scenario 3**: We generated X_1 and X_2 from the t-distribution, with 50 observations per class. The t-distribution has a similar shape to the normal distribution, but it has a tendency to yield more extreme points—that is, more points that are far from the mean. In this setting, the decision boundary was still linear, and so fit into the logistic regression framework. The set-up violated the assumptions of LDA, since the observations were not drawn from a normal distribution. The right-hand panel of Figure 4.10 shows that logistic regression outperformed LDA, though both methods were superior to the other approaches. In particular, the QDA results deteriorated considerably as a consequence of non-normality.



- **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. This setup corresponded to the QDA assumption, and resulted in quadratic decision boundaries. The left-hand panel of Figure 4.11 shows that QDA outperformed all of the other approaches.

- **Scenario 5**: Within each class, the observations were generated from a normal distribution with uncorrelated predictors. However, the responses were sampled from the logistic function using X_1^2, X_2^2, and $X1 * X2$ as predictors. Consequently, there is a quadratic decision boundary. The center panel of Figure 4.11 indicates that QDA once again performed best, followed closely by KNN-CV. The linear methods had poor performance.

- **Scenario 6**: Details are as in the previous scenario, but the responses were sampled from a more complicated non-linear function. As a result, even the quadratic decision boundaries of QDA could not adequately model the data. The right-hand panel of Figure 4.11 shows that QDA gave slightly better results than the linear methods, while the much more flexible KNN-CV method gave the best results. But KNN with K = 1 gave the worst results out of all methods. This highlights the fact that even when the data exhibits a complex non-linear relationship, a non-parametric method such as KNN can still give poor results if the level of smoothness is not chosen correctly.



This shows that no one method can dominate the others in every situation.

- When the true decision boundaries are linear, then the LDA and logistic regression approaches will tend to perform well.
- When the boundaries are moderately non-linear, QDA may give better results.
- 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.