3 LDA "linear discriminant analysis"

Logistic regression involves ~~direction~~ directly modeling $P(Y = k|X = x)$ using the logistic function for the case of two response classes. We now consider a less direct approach.

Idea:

$P(X=x|Y=y)$: Model the distribution of the predictors X separately in each of the response classes (given Y) and then use Bayes theorem to flip these and get estimates for $P(Y=k|X=x)$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Why do we need another method when we have logistic regression?

1. When classes are well-separated, the parameter estimates for logistic regression are suprisingly unstable.

2. If n is small and the distribution of predictors is approximately normal in each class, LDA is more stable than logistic regression.

3. We might have more than 2 response classes.

3.1 Bayes' Theorem for Classification

Suppose we wish to classify an observation into one of K classes, where $K \geq 2$.

Categorical Y can take on K possible distinct and unordered values.

notation

π_k - overall or "prior" probability that a randomly chosen observation comes from the k^{th} class.

$f_k(x) = P(X=x \mid Y=k)$ (discrete)

probability that X takes value x given Y of class k.

density function of X for an observation that comes from class k.

(CTS) probability that X falls in a small region around x given $Y=k$.

$$\overset{A}{P(Y=k}|\overset{B}{X=x)} = \frac{\overset{P(B|A)}{\overset{P(X=x|Y=k)}{f_k(x)}}\ \overset{P(A)}{\overset{P(Y=k)}{\pi_k}}}{\underbrace{\sum_{\ell=1}^{K} \underbrace{\pi_\ell f_\ell(x)}_{P(X=x,\,Y=k)}}_{P(X=x)\ P(B)}}$$ (Bayes Theorem)

We will use the abbreviation $p_k(x)$ as before.

> Called the "posterior probability" that an observation comes from k^{th} class given $X=x$.

In general, estimating π_k is easy if we have a random sample of Y's from the population.

Compute the fraction of training observations that come from the k^{th} class.

Estimating $f_k(x)$ is more difficult unless we assume some particular forms.

If we can estimate $f_k(x)$ we can devlop a classifier that is close to the "best" classifier (more later).

3.2 p = 1

p=1.

Let's (for now) assume we only have 1 predictor. We would like to obtain an estimate for $f_k(x)$ that we can plug into our formula to estimate $p_k(x)$. We will then classify an observation to the class for which $\hat{p}_k(x)$ is greatest.

↳ $\frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$

estimating the Bayes classifier.

assign to class w/ $p_k(x)$ highest is called the "Bayes classifier" and is known to be the optimal solution, i.e. we can do no better!

Suppose we assume that $f_k(x)$ is normal. In the one-dimensional setting, the normal density takes the form

$$f_K(x) = \frac{1}{\sqrt{2\pi\sigma_K^2}} \exp\left[-\frac{1}{2\sigma_K^2}(x-\mu_K)^2\right]$$

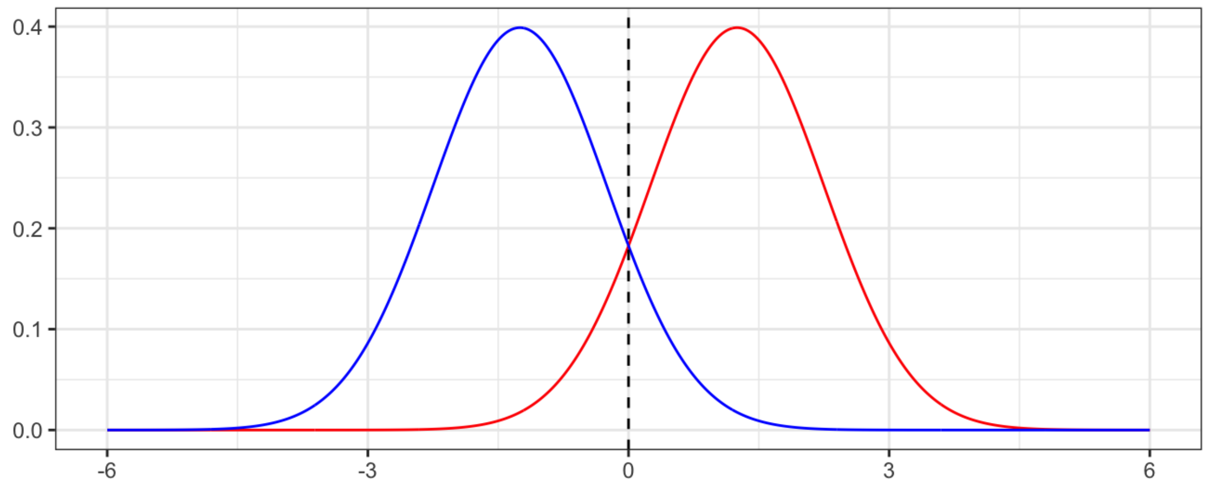
σ_K^2 and μ_K variance and mean parameters for Kth class.

Let's also (for now) assume $\sigma_1^2 = \dots = \sigma_K^2 = \sigma^2$ (shared variance term).

Plugging this into our formula to estimate $p_k(x)$,

We then assign an observation $X = x$ to the class which makes $p_k(x)$ tthe largest. This is equivalent to

Example 3.1 Let $K = 2$ and $\pi_1 = \pi_2$. When does the Bayes classifier assign an observation to class 1?

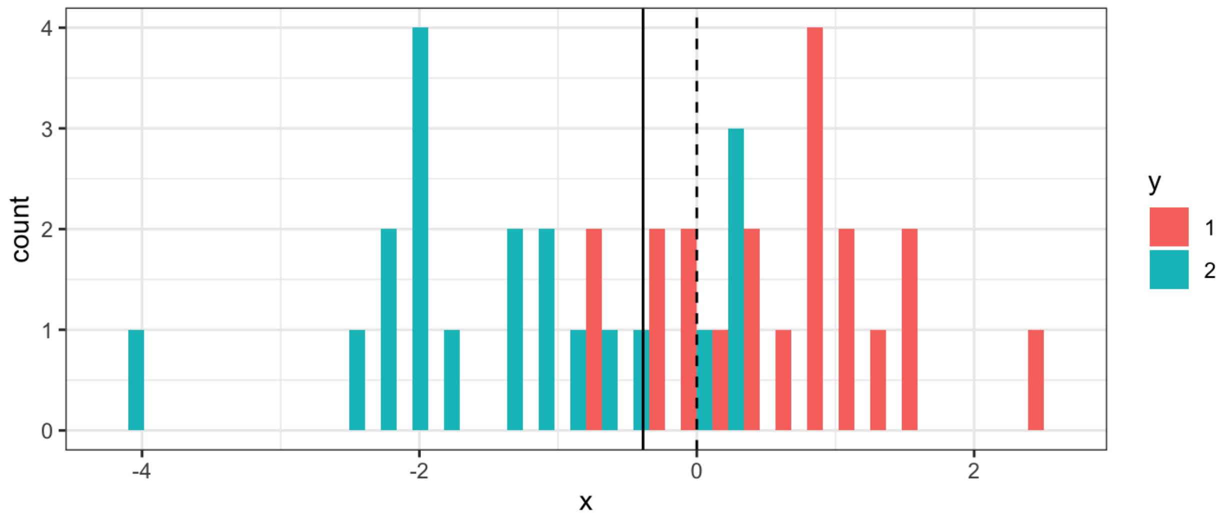


In practice, even if we are certain of our assumption that X is drawn from a Gaussian distribution within each class, we still have to estimate the parameters $\mu_1, \dots, \mu_K, \pi_1, \dots, \pi_K, \sigma^2$.

The *linear discriminant analysis* (LDA) method approximated the Bayes classifier by plugging estimates in for π_k, μ_k, σ^2.

Sometimes we have knowledge of class membership probabilities $\pi_1, \dots, \pi_K$ that can be used directly. If we do not, LDA estimates π_k using the proportion of training observations that belong to the kth class.

The LDA classifier assignes an observation $X = x$ to the class with the highest value of




```
##    pred
## y       1     2
##   1 18966  1034
##   2  3855 16145
```

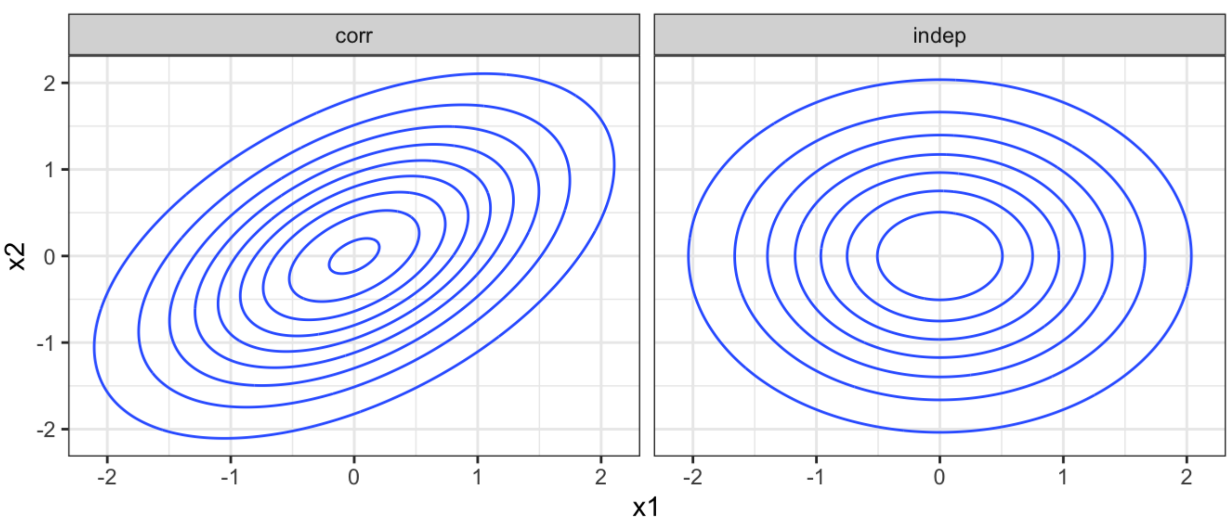
The LDA test error rate is approximately 12.22% while the Bayes classifier error rate is approximately 10.52%.

The LDA classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance σ^2 and plugging estimates for these parameters into the Bayes classifier.

3.3 p > 1

We now extend the LDA classifier to the case of multiple predictors. We will assume

Formally the multivariate Gaussian density is defined as



In the case of $p > 1$ predictors, the LDA classifier assumes the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.

Plugging in the density function for the kth class, results in a Bayes classifier

Once again, we need to estimate the unknown parameters $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \pi_1, \dots, \pi_K, \boldsymbol{\Sigma}$.

To classify a new value $X = x$, LDA plugs in estimates into $\delta_k(x)$ and chooses the class which maximized this value.

Let's perform LDA on the `Default` data set to predict if an individual will default on their CC payment based on balance and student status.

```r
library(MASS) # package containing lda function
lda_fit <- lda(default ~ student + balance, data = Default)
lda_fit
```

```
## Call:
## lda(default ~ student + balance, data = Default)
##
## Prior probabilities of groups:
##     No    Yes
## 0.9667 0.0333
##
## Group means:
##     studentYes   balance
## No   0.2914037  803.9438
## Yes  0.3813814 1747.8217
##
## Coefficients of linear discriminants:
##                      LD1
## studentYes -0.249059498
## balance     0.002244397
```

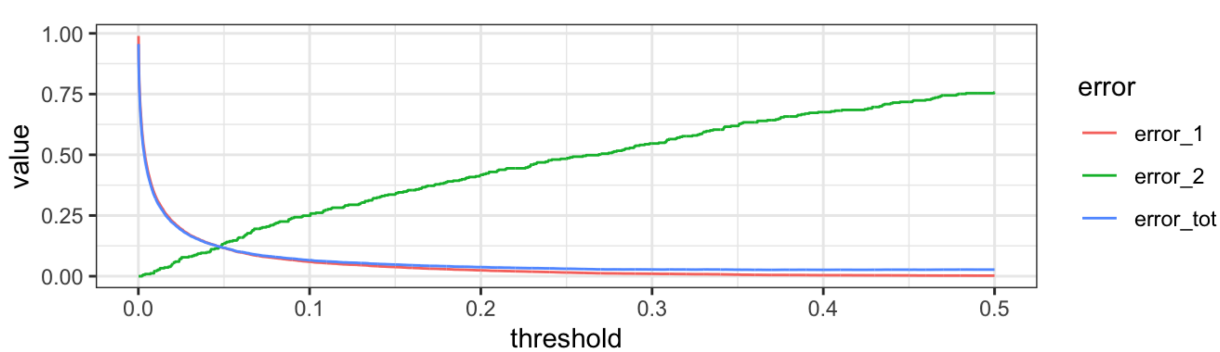
```r
# training data confusion matrix
table(predict(lda_fit)$class, Default$default)
```

```
##
##         No  Yes
##   No  9644  252
##   Yes   23   81
```

Why does the LDA classifier do such a poor job of classifying the customers who default?

```r
table(predict(lda_fit)$posterior[, "Yes"] > 0.2, Default$default)
```

```
##
##           No  Yes
##   FALSE 9432  138
##   TRUE   235  195
```



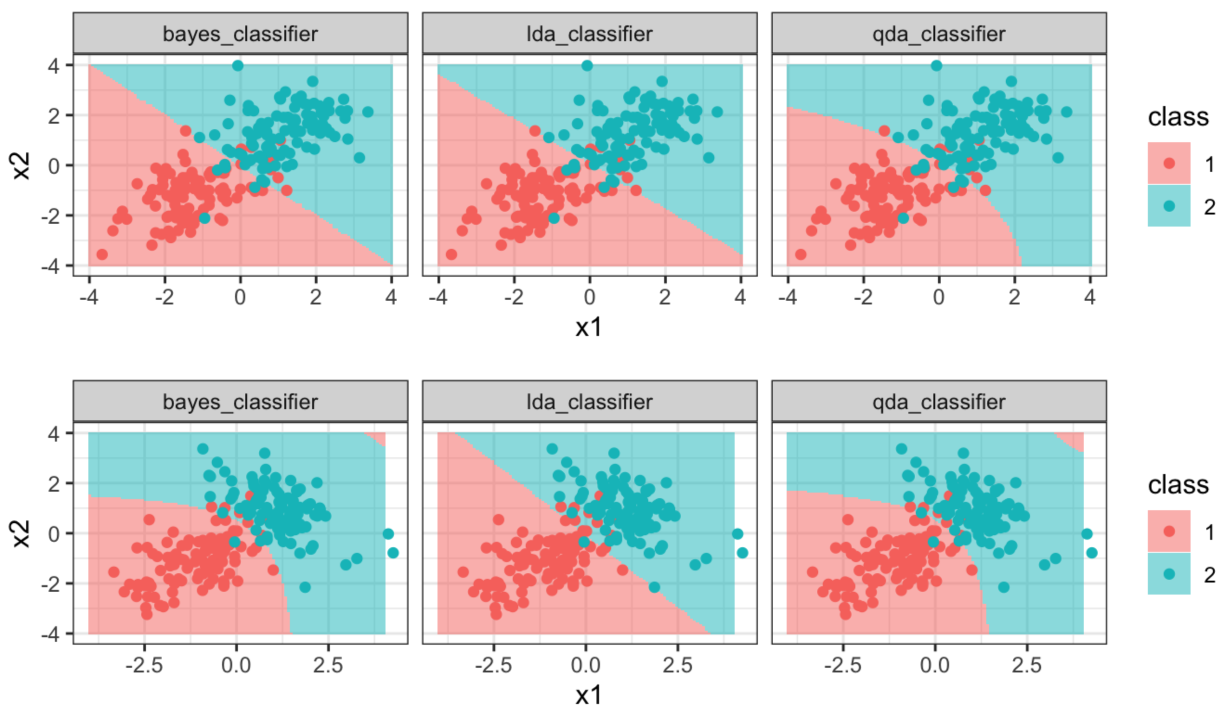
3.4 QDA

LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class-specific mean vector and a common covariance matrix across all K classes.

Quadratic Discriminant Analysis (QDA) also assumes the observations within each class are drawn from a multivariate Gaussian distribution with a class-specific mean vector but now each class has its own covariance matrix.

Under this assumption, the Bayes classifier assignes observation $X = x$ to class k for whichever k maximizes

When would we prefer QDA over LDA?

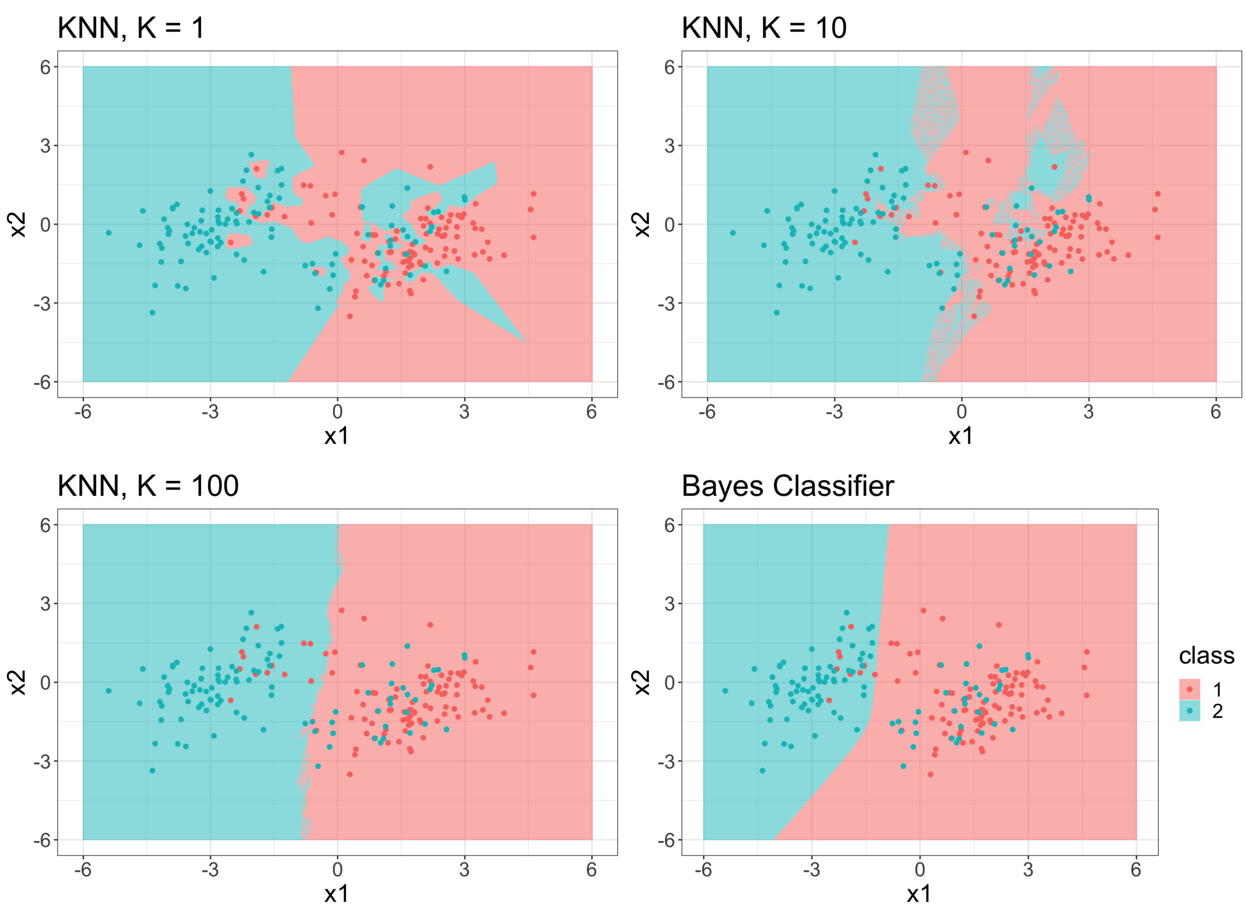


4 KNN

Another method we can use to estimate $P(Y = k|X = x)$ (and thus estimate the Bayes classifier) is through the use of K-nearest neighbors.

The KNN classifier first identifies the K points in the training data that are closest to the test data point $X = x$, called $\mathcal{N}(x)$.

Just as with regression tasks, the choice of K (neighborhood size) has a drastic effect on the KNN classifier obtained.



5 Comparison

LDA vs. Logistic Regression

(LDA & Logistic Regression) vs. KNN

QDA