Advanced Topics in Statistics

~ Classification - LDA, QDA ~

3

UNIVERSITY OF EXETER

Introduction.

Logistic regression involves directly modelling the distribution of the classes given the predictor, that is $P(Y = k|\boldsymbol{X} = \boldsymbol{x})$, using the logistic function.

An alternative, less direct approach is modelling the distribution of the predictors $\boldsymbol{X}$ separately in each class (i.e. given Y), and using Bayes theorem to flip the conditioning around.

Under certain assumptions this indirect method is called linear discriminant analysis.

Why do we need another method when we can use logistic regression for classification?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors $\boldsymbol{X}$ is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes.

Using Bayes theorem for classification.

Suppose that the response variable Y can take on K possible distinct and unordered values, where $K \geq 2$.

- Let π_k represent the prior probability that a randomly chosen observation comes from the kth class.
- Let $f_k(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|Y = k)$ denote the density function of $\boldsymbol{X}$ for an observation that comes from the kth class.

  In other words $f_k(\boldsymbol{x})$ is relatively large if there is a high probability that an observation in the kth class has $\boldsymbol{X} \approx \boldsymbol{x}$, and $f_k(\boldsymbol{x})$ is small if it is very unlikely that an observation in the kth class has $\boldsymbol{X} \approx \boldsymbol{x}$.
- Then using Bayes theorem gives

$$p_k(\boldsymbol{x}) := P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x})}.$$

  So instead of calculating the probability of belonging to a class directly, we plug the estimates of π_k and $f_k(\boldsymbol{x})$ into the above equation.
- We refer to $p_k(x)$ as the posterior probability that an observation $\boldsymbol{X} = \boldsymbol{x}$ belongs to the kth class.

LDA ASSUMPTIONS.

In general, estimating π_k is easy: if we have a random sample of Ys from the population we compute the fraction of the training observations that belong to the kth class.

Estimating $f_k(\boldsymbol{X})$ tends to be more challenging, unless we assume some simple forms for these densities.

The linear discriminant analysis makes the following assumptions, which allow us to find $f_k(\boldsymbol{X})$ (given that the assumptions hold):

- For a single predictor: $f_k(X)$ has a normal distribution with class specific mean μ_k, and variance σ_k^2.

  But the variances are shared across all K classes, thus

  $$\sigma^2 := \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2.$$

- When we have more than one predictors, this translates to: $f_k(\boldsymbol{X})$ has a multivariate normal distribution with class-specific mean vector $\boldsymbol{\mu}_k$ and shared covariance matrix $\boldsymbol{\Sigma}$.

Once we have π_k and $f_k(\boldsymbol{X})$, we can estimate $p_k(\boldsymbol{x})$ using Bayes theorem. LDA then assigns the observation $\boldsymbol{X} = \boldsymbol{x}$ to the class for which $p_k(\boldsymbol{x})$ is the largest. By doing so LDA approximates the Bayes classifier.

Bayes classifier under the LDA assumptions, $p = 1$.

In one dimension, the normal distribution has probability density function

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right).$$

Then using Bayes theorem (and the equality of the variances assumption) we get

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)},$$

where π_k denotes the prior probability that an observation belongs to the kth class, while π (without the subscript) is the mathematical constant 3.14159....

The Bayes classifier assigns an observation $X = x$ to the class for which this is largest.

After taking the log of $p_k(x)$, we can see that this is equivalent to assigning the observation to the class for which

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

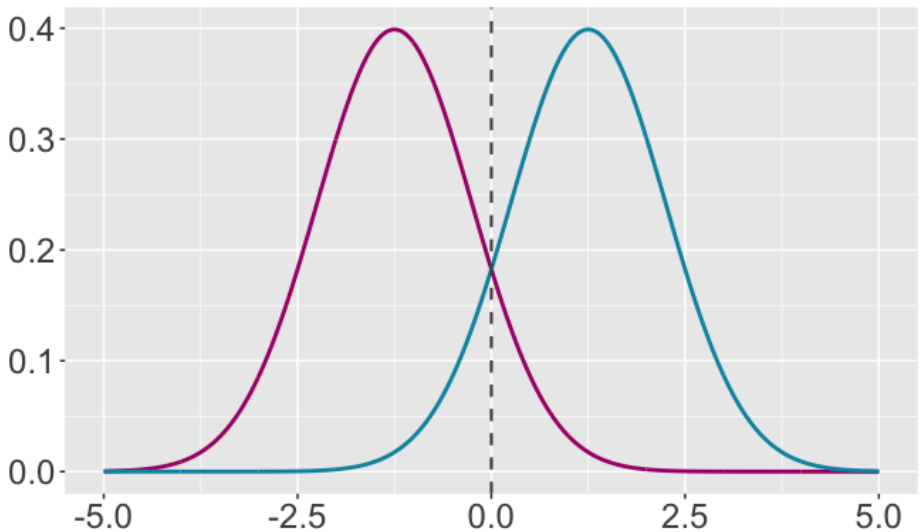
is the largest.

Bayes decision boundary example.

If $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1 if $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, and to class 2 otherwise.

The Bayes decision boundary then corresponds to the points where

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.$$



- In this case we have two normal density functions $f_1(x)$ and $f_2(x)$ that represent two distinct classes.
- The two density functions overlap, so there is some uncertainty about the class to which an observation with an unknown class belongs.

The idea is to choose the class whose density is highest at the given location.

(The dashed vertical line represents Bayes' decision boundary).

LDA ESTIMATES.

In real life situations we are not able to calculate the Bayes classifier.

Instead LDA approximates the Bayes classifier by plugging estimates for μ_k, σ^2 and π_k into the expression of $\delta_k(x)$. In particular, the following estimates are used.

- The estimate of the class specific mean μ_k is the average of all the training observations from the kth class, that is

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i,$$

  where n_k denotes the number of observations in the kth class.
- The estimate of the shared variance σ^2 can be seen as a weighted average of the sample variances for each of the K classes, that is

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2.$$

- In the absence of any additional information, LDA estimates π_k using the proportion of the training observations that belong to the kth class

$$\hat{\pi}_k = \frac{n_k}{n}.$$

LINEAR DISCRIMINANT ANALYSIS.

After plugging the previous estimates for π_k, μ_k and σ^2 into the expression for $\delta_k(x)$ we get that LDA assigns the observation to the class for which

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}_k} - \frac{\hat{\mu}^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

is the largest.

We call $\delta_k(x), k = 1, \ldots, K$ the discriminant functions.

Linear refers to the fact that the discriminant functions $\hat{\delta}_k(x)$ are linear functions of x (and nothing more complex than that!).

Now assume we have more than one predictor, that is $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is drawn from a multivariate Gaussian (multivariate normal) distribution, with a class-specific mean vector and a common covariance matrix.

Then after some algebra we can see that LDA assigns the observation $\boldsymbol{X} = \boldsymbol{x}$ to the class for which

$$\hat{\delta}_k(\boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\hat{\pi}_k)$$

is the largest. This is the vector/matrix version of the previous expression.

LDA SUMMARY.

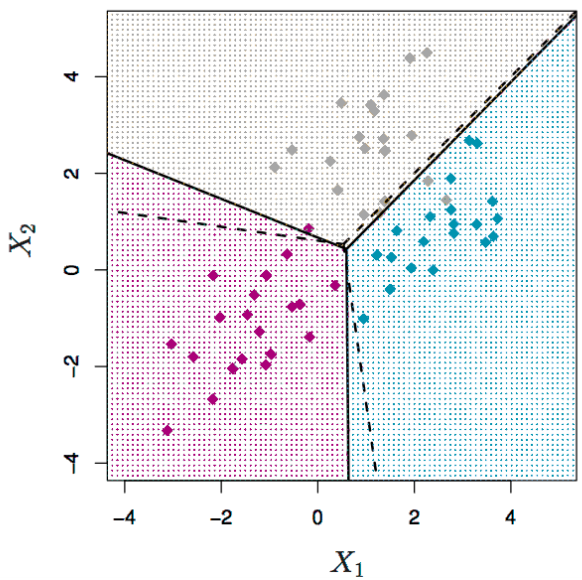
Overall LDA involves the determination of linear equations for each class in the form of

$$D = \nu_1 X_1 + \nu_2 X_2 + \cdots + \nu_p X_p + a,$$

where D is the discriminant function, $\boldsymbol{\nu}$ is the discriminant coefficient (or weight for the variable), $\boldsymbol{X}$ is the variable, and a is a constant.

LDA then picks the class with the largest associated D value.

Thus the LDA decision rule depends on $\boldsymbol{x}$ through a linear combination of its elements, which once again shows the linear nature of the LDA classifier.



In this simulated example we have two predictors ($p = 2$), and three classes.

Twenty observations were generated from each class.

The solid lines are the Bayes' boundaries.

The dashed linear are LDA boundaries.

Running LDA on credit card default data.

Recall the credit card default data example, where we tried to predict whether an individual will default on the basis of credit card balance and student status.

Here we will use LDA instead of logistic regression to classify the observations, which gives the following confusion matrix

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- LDA makes 252+23 mistakes on 10000 predictions (2.75% misclassification error rate), which sounds like a low error rate, but we have to keep in mind that training error rates are usually lower than test error rates, which are the real quantity of interest.
- And since only 3.33% of the individuals in the training sample defaulted, even the trivial null classifier which always predicts 'no default' will result in an error rate of 3.33%, not much higher than the LDA training set error rate.
- Also notice that LDA miss-predicts 252/333 = 75.5% of defaulters!

Comment on the model fit.

In the credit card example, the basic LDA has a 24.3% sensitivity and 99.8% specificity.

The credit card company however might want to avoid incorrectly classifying an individual who will default, whereas incorrectly classifying an individual who will not default is less problematic.

It is possible to modify the LDA algorithm so that it better meets the credit card company's needs.

Since LDA is an approximation of the Bayes classifier, which uses $p^* = 0.5$ as a threshold in a binary classification problem, by extension LDA also uses $p^* = 0.5$ as a threshold. That is, it assigns an observation to the default class if

$$P(\texttt{default=Yes}|\boldsymbol{X} = \boldsymbol{x}) > 0.5.$$

We could however lower this threshold if we are concerned about incorrectly predicting the default status for individuals who default.

For example, choosing $p^* = 0.2$, that is assigning an observation to the default class if

$$P(\texttt{default=Yes}|\boldsymbol{X} = \boldsymbol{x}) > 0.2$$

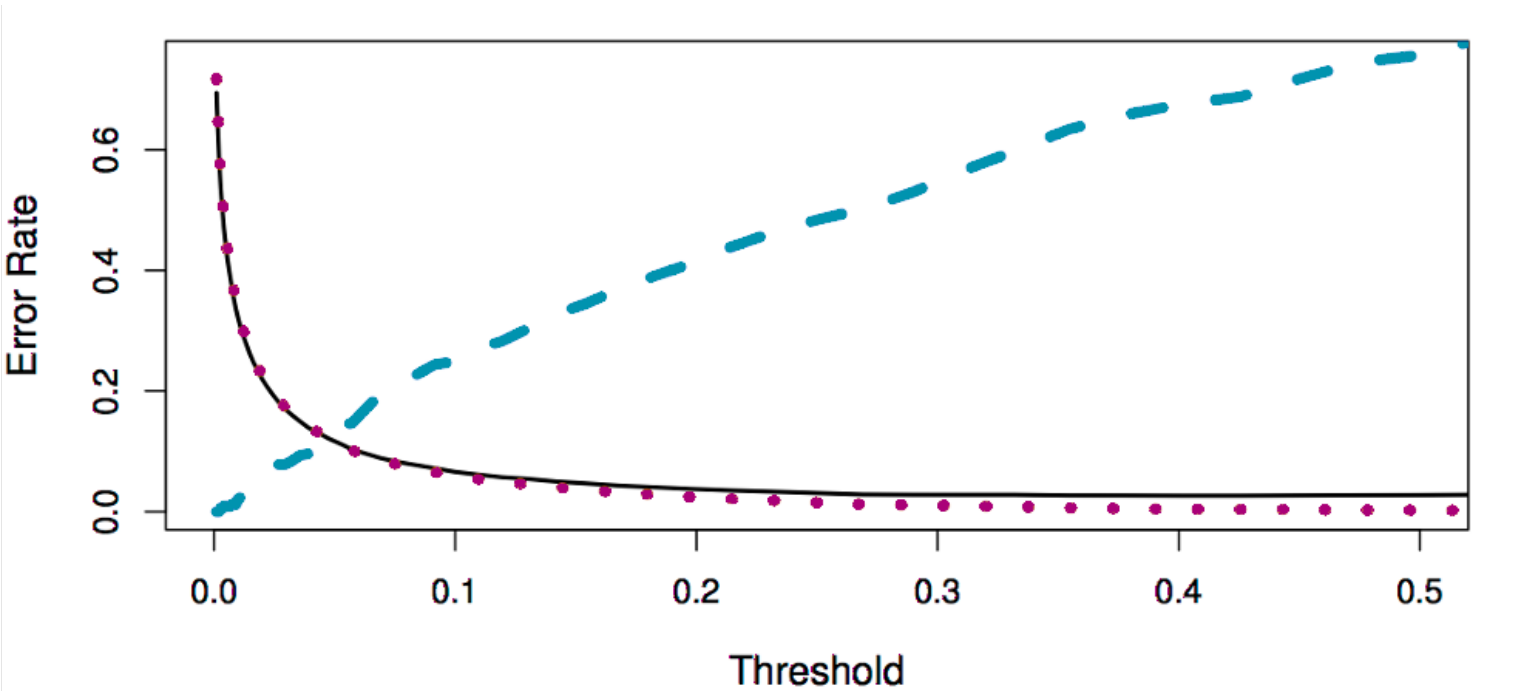
would result in a 3.73% misclassification error rate, but we would only mispredict 41.4% defaulters.

USING DIFFERENT THRESHOLDS.

The figure below shows the trade-off that results from modifying the threshold value for the posterior probability of default.

The following error rates are shown as a function of the threshold value:

- **Black** solid: overall error rate (lowest at $p^* = 0.5$).
- **Blue** dashed: Fraction of defaulters missed.
- **Pink** dotted: non defaulters incorrectly classified.



Quadratic Discriminant Analysis.

As we have seen LDA assumed that every class has the same variance/covariance.

However, LDA may perform poorly if this assumption is far from true.

Quadratic discriminant analysis (QDA) works identically as LDA except that it estimates separate variances/ covariance for each class.

QDA keeps the normality assumption, that is it assumes that the observations from each class are drawn from a Gaussian distribution, but allows for class-specific covariance matrices.

Under these assumptions the Bayes classifier assigns an observation $\boldsymbol{X} = \boldsymbol{x}$ to the class for which

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k) \\ &= -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \log(\pi_k)\end{aligned}$$

is the largest.

QDA involves plugging estimates for $\boldsymbol{\Sigma}_k$, $\boldsymbol{\mu}_k$ and π_k into the above expression.

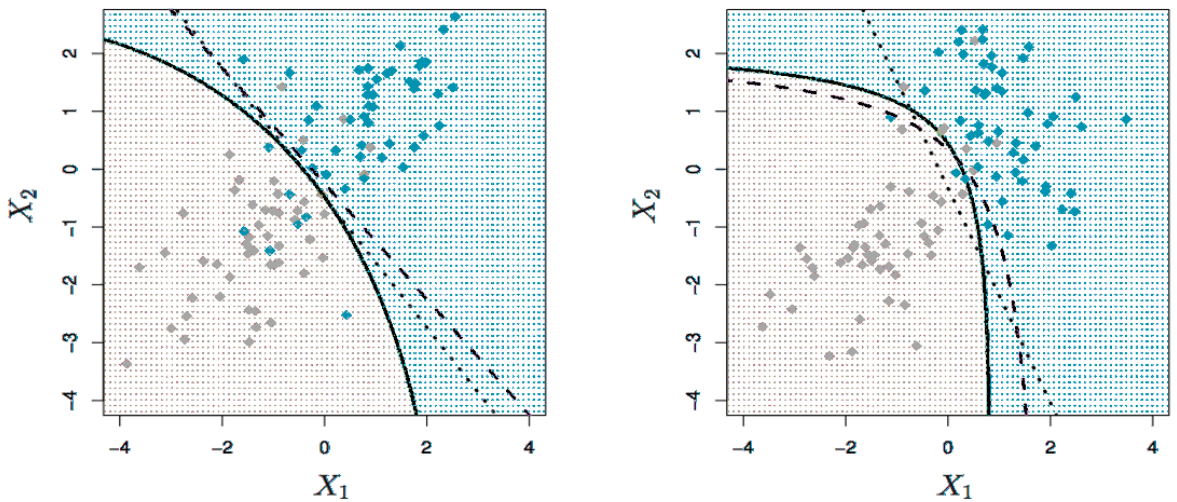
Unlike the LDA case, the discriminant functions $\hat{\delta}_k(\boldsymbol{x})$ are now a quadratic functions of the quantity $\boldsymbol{x}$, which results in a quadratic decision boundary.

LDA vs QDA.

QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate these variances (assuming class-specific variances/covariances means that we have a lot more parameters to estimate from the same amount of data).

LDA will work best when the variances are similar among classes or we don't have enough data to accurately estimate the variances.

In general, LDA tends to have lower variance, but higher bias.



Dotted:
LDA boundary.

Dashed:
Bayes' boundary.

Solid:
QDA boundary.

Left: variances of the classes are equal (LDA is better fit).

Right: variances of the classes are not equal (QDA is better fit).

Comparison of different classification methods.

So far we have considered four classification methods: K-nearest neighbours, logistic regression, LDA, QDA.

In what scenarios would one dominate the others?

Logistic regression vs LDA

- Both logistic regression and LDA methods produce linear decision boundaries (in the default setting at least).
- The only difference is that the coefficients of the logistic regression are estimated using maximum likelihood, while the weights of the LDA method are computed using the estimated means and variance from a normal distribution.
- They often give similar results, but not always.

  LDA assumes that the observations are drawn from a Gaussian distribution with a common covariance matrix in each class, and so provide improvements over logistic regression when these the assumptions are met.

  But logistic regression can outperform LDA if the Gaussian assumptions are not met.

Comparison of different classification methods.

KNN vs LDA, Logistic regression

- KNN is a completely non-parametric approach, no assumptions are made about the shape of the decision boundary.
- Hence it can dominate LDA and logistic regression when the decision boundary is highly non-linear. (But KNN does not tell us which predictors are important).
- We also have to choose the level of smoothness when doing KNN classification, and so we have to keep the dangers of overfitting in mind.

QDA vs KNN and LDA, logistic regression

- QDA serves as a compromise between the non-parametric KNN methods and the linear LDA and logistic regression approaches.
- The quadratic decision boundary can be more appropriate than a linear one in many scenarios. In other cases we need the higher flexibility of KNN.
- QDA can perform better than KNN in the presence of a limited number of training observations, because it does make some assumptions about the form of the decision boundary.