Contents

1 Linear discriminant analysis

1.1 LDA approach

- Logistic regression involves *directly* modeling $Pr(Y = k|X = x)$ using the logistic function, e.g.

$$Pr(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

  for the case of two response classes.

- Here the approach is to model the distribution of X in each of the classes separately (i.e. given Y), and then use **Bayes' theorem** to *flip* these around into estimates for $Pr(Y|X)$.
- When we use *normal* (Gaussian) distributions for each class, this leads to **linear** or **quadratic** discriminant analysis.
- However, this approach is quite general, and other distributions can be used as well. We will focus on normal distributions.

Why a new approach?

- When the distribution of X for each class is normal, it turns out that the model is very similar in form to logistic regression.
  - LDA estimates are *more stable* than those from logistic regression
    - when the classes are well-separated,
    - If n is small and the distribution of the predictors X is approximately normal in each of the classes
- linear discriminant analysis is popular when we have *more than two response classes.*

1.2 Bayes theorem for classification

Bayes theorem

$$Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k) \cdot Pr(Y = k)}{Pr(X = x)}$$

Bayes theorem for DA

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{1}$$

- $\pi_k = Pr(Y = k)$ is the marginal or **prior** probability for class k.
- $f_k(x) = Pr(X = x|Y = k)$ is the **density** for X in class k.
  - Here we will use normal densities for these, separately in each class.
- $p_k(x) = Pr(Y = k|X = x)$ is the **posterior** probability that an observation X belongs to the class k.

- In general, estimating π_k is easy if we have a random sample of Ys: we simply compute the fraction of the training observations that belong to the kth class.
- Estimating $f_k(x)$ is more challenging, unless we assume some simple forms for these densities.

- The **Bayes classifier** which classifies an observation to the class for which $p_k(x)$ is largest, has the lowest possible (test) error rate out of all classifiers.
- Therefore, if we can find a way to estimate $f_k(x)$, then we can *develop a classifier that approximates the Bayes classifier.*

Recall: Test error rate

Suppose that we seek to estimate f on the basis of training observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. In a classification setting, $y_1, \ldots, y_n$ are qualitative.

The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the *training* **error rate**

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

that is the fraction of incorrect classifications.

The *test* **error rate** associated with a set of test observations of the form (x_*, y_*) is given by

$$\mathrm{E}_X(I(y_* \neq \hat{y}_*))$$

the average error rate on test observations over the distribution of X.

Bayes classifier:

assigns each observation to the most likely class, given its predictor values, i.e. assign an observation with predictor vector x to the class j for which

$$Pr(Y = j \,|\, x) \quad \text{is largest}$$

- $Pr(Y = k|X = x)$ is a conditional probability.
- In a two-class problem (1/2), when does the Bayes classifier predict class one*?*
- The Bayes classifier produces *the lowest possible (test) error rate*, called the **Bayes error rate**

$$1 - E_X\left(\max_j P(Y = j|X)\right).$$

(It is analogous to the irreducible error discussed earlier.)

LDA focuses on estimating $f_k(X)$ by which together with the prior π_k provides the posterior probability $p_k(X)$, thus it provides a classifier which approximates the Bayes classifier.

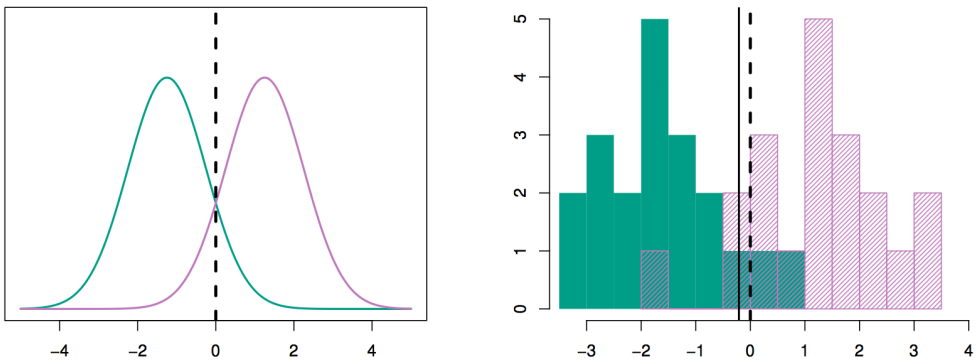


Figure 1: LDA when $p = 1$. Histograms of 20 obs for each class are shown.

1.3 LDA when $p = 1$

The Gaussian density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

* We will *assume* that all the $\sigma_k = \sigma$ are the same.

Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = Pr(Y = k|X = x)$

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

Happily, there are simplifications and cancellations.

1.3.1 Discriminant functions

To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest.

Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest **discriminant score**:

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

Note that $\delta_k(x)$ is a *linear* function of x.

If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$ then one can see that the **decision boundary** is at

$$x = \frac{\mu_1 + \mu_2}{2}. \tag{3}$$

Illustration: $p = 1$, $K = 2$, $\pi_1 = \pi_2$

Example with $\mu_1 = -1.5$, $\mu_2 = 1.5$, and $\sigma^2 = 1$.

- The dashed line is known as the **Bayes decision boundary**.
  - Were it known, it would yield the fewest misclassification errors, among all possible classifiers.

- Typically we don't know the model parameters; we just have the training data. In that case we simply estimate the parameters and plug them into the rule.

Parameters are estimated as follows:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:\, y_i = k} x_i \tag{4}$$

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

And, in the absence of any additional information,

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

The LDA classifier approximates the Bayes classifier by plugging the estimates (4) into the discriminant function (2), and assigns an observation $X = x$ to the class k for which

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

is largest.

Analogously, the decision boundary (3) is estimated as

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}.$$

In the illustration above, the LDA decision boundary is shown by the solid black line.

How well does the LDA classifier perform on this data?

The Bayes error rate is 10.6%, the LDA test error rate is 11.1%.

1.4 LDA when $p > 1$

$X = (X_1, \ldots, X_p) \sim N(\mu, \Sigma)$, density: $f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$

What are the differences between the two MVN distributions in figure?

LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \Sigma)$.}

Plugging the density function for the kth class, $f_k(X = x)$, into (1), the discriminant function is

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$$

Despite its complex form,

$$\delta_k(x) = c_{k0} + c_{k1} x_1 + c_{k2} x_2 + \ldots + c_{kp} x_p$$

is a linear function.

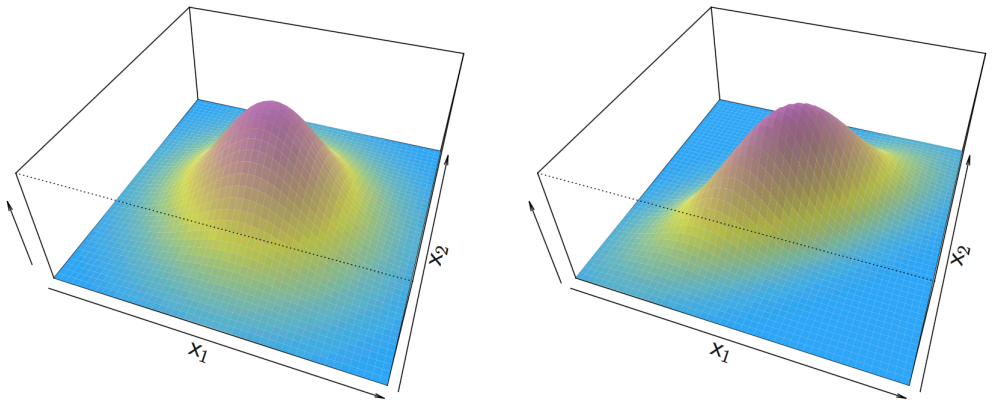


Figure 2: Two examples of 2-variate Normal.

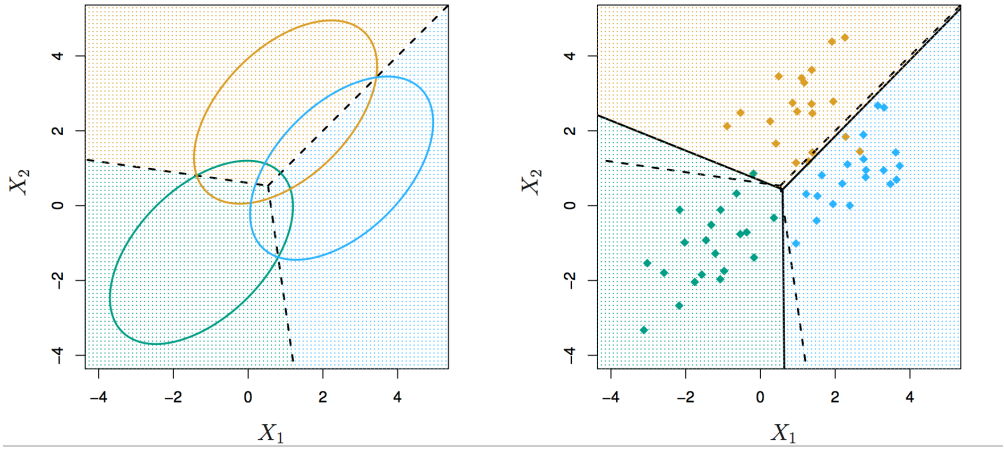


Figure 3: LDA when $p = 2$. Points of 20 obs. for each class are shown.

Illustration: $p = 2$, $K = 3$, $\pi_1 = \pi_2 = \pi_3$

The three ellipses represent regions that contain 95% of the probability for each of the three classes.

The dashed lines are the Bayes decision boundaries. (Were they known, they would yield the fewest misclassification errors, among all possible classifiers.)

LDA classifier estimates parameters by formulas similar to those seen for the case $p = 1$.

Plug-in these estimates into the discriminant function.

Solid black lines show the LDA decision boundaries.

The test error rates for the Bayes and LDA classifiers are 0.0746 and 0.0770, respectively.

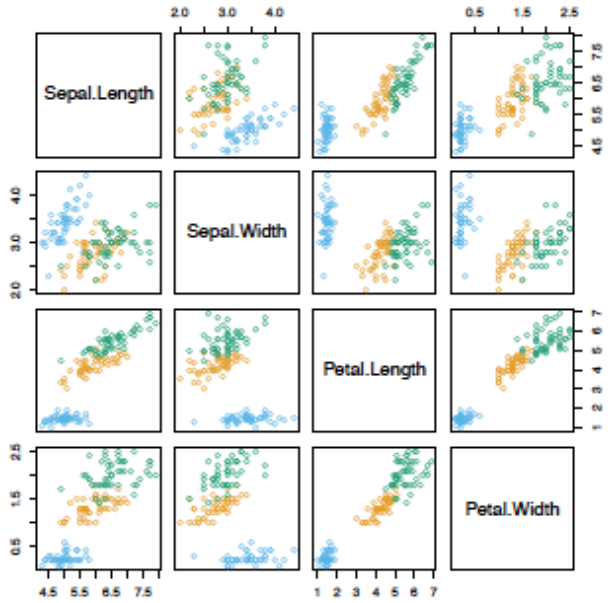
From $\delta_k(x)$ to probabilities

Once we have estimates $\hat{\delta}_k(x)$ we can turn these into estimates for class probabilities

$$\hat{Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\hat{Pr}(Y = k|X = x)$ is largest.

Fisher's Iris Data



4 variables, 3 species (Setosa (blue), Versicolor (orange), Virginica (green)), 50 samples/class.

LDA classifies all but 3 of the 150 training samples correctly.

1.5 Example

LDA on Credit Data: the **confusion matrix**

Predicted (by row) / True (by col) Default status

	No	Yes	Total
No	9,644	252	9,896
Yes	23	81	104
Total	9,667	333	10,000

What is the error rate?

(23+252)/10000 errors, a 2.75% misclassification rate!

Some caveats:

- This is training error, and we may be **overfitting**.
  - Not a big concern here since $n = 10,000$ and $p = 3$.
- If we classified by a *Null classifier*—always to class `No` in this case—how many errors would we make?

333/10000 errors, or only 3.33%!

- Of the true `No`'s, we make 23/9667 = 0.2% errors; of the true `Yes`'s, we make 252/333 = 75.7% errors!!!

Types of errors

- **False positive rate** The fraction of negative examples that are classified as positive – 0.2% in example
- **False negative rate**
  The fraction of positive examples that are classified as negative - 75.7% in example

In medicine and biology,

- the *complement to FP rate* is called **specificity**,
- the *complement to FN rate* is called **sensitivity**.

In our example, has LDA a low sensitivity or specificity? It has a low sensitivity: 24.3%

Is the default threshold adequate?

We produced this table by classifying to class `Yes` if

$$\hat{Pr}(\texttt{Default} = \texttt{Yes}|\texttt{Balance}, \texttt{Student}) \geq 0.5$$

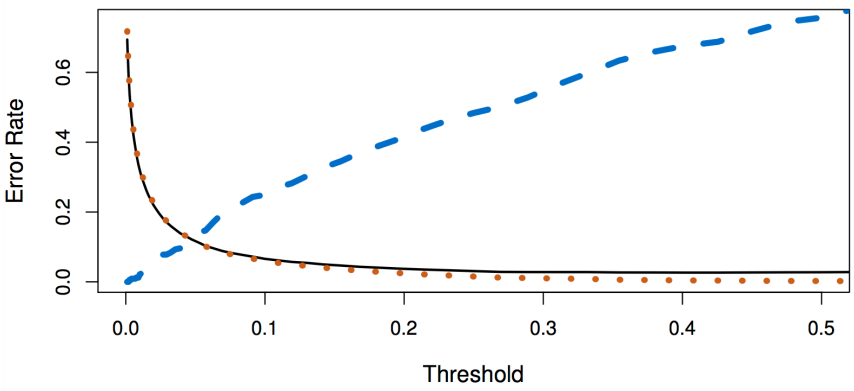
We can change the two error rates by changing the threshold from 0.5 to some other value in $[0, 1]$:

$$\hat{Pr}(\texttt{Default} = \texttt{Yes}|\texttt{Balance}, \texttt{Student}) \geq \text{threshold}$$

and vary *threshold*.

In order to reduce the FN rate, should we lower or raise the threshold ?

Varying the threshold

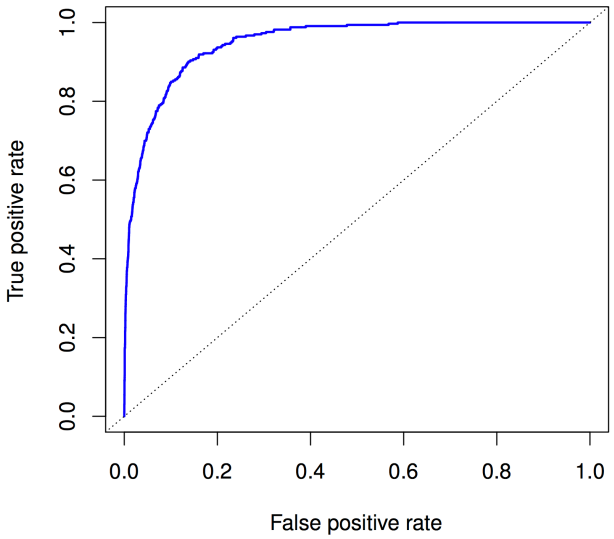


The lines represent Overall Error, False Positive (red), False Negative (blue)

In order to reduce the FN rate, we may want to reduce the threshold to 0.1 or less.

Such a decision must be based on **domain knowledge**.

ROC Curve



The **ROC plot** displays both errors simultaneously.

What is the TP rate?

Sometimes we use the **AUC** or area under the curve to summarize the overall performance. Higher AUC is good.

1.6 Other DAs

Other forms of Discriminant Analysis

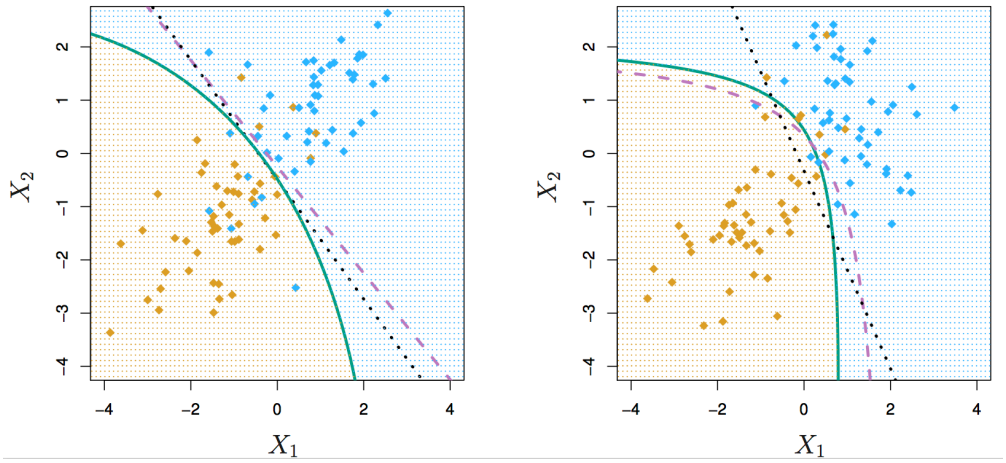
$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

When $f_k(x)$ are Gaussian densities, with the same covariance matrix Σ in each class, this leads to linear discriminant analysis.

By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different Σ_k in each class, we get **quadratic discriminant analysis**.
- With $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$ (conditional independence model) in each class we get **naive Bayes**.
  - For Gaussian this means the Σ_k are diagonal.
- Many other forms, by proposing specific density models for $f_k(x)$, including nonparametric approaches.

1.6.1 Quadratic Discriminant Analysis



Bayes, LDA and QDA at comparison when $\Sigma_1 = \Sigma_2$ (left) and $\Sigma_1 \neq \Sigma_2$ (right).

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

Because the Σ_k are different, the quadratic terms matter.

What is more flexible: LDA or QDA?

Which consequences derive from choosing one or the other on variance and bias of the classifier?

Logistic regression vs LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1x_1 + \ldots + c_px_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated (ML vs moment estimation).

- Logistic regression uses the conditional likelihood based on $Pr(Y|X)$ (known as **discriminative learning**).
- LDA uses the full likelihood based on $Pr(Y, X)$ (known as **generative learning**). Despite these differences, in practice the results are often very similar.

Logistic regression can also fit quadratic boundaries like QDA, by explicitly including quadratic terms in the model.

Summary

- Logistic regression is very popular for classification, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Naive Bayes is useful when p is very large.
- Next we present the k-nearest neighbour classifier, a nonparametric method useful when the decision boundary is very irregular.