Lecture 3: Classification

Can Yang

Department of Mathematics
The Hong Kong University of Science and Technology

Most of the materials here are from Chapter 4 of Introduction to Statistical learning by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

1. 4.1. Overview
2. 4.2. Relation with regression
3. 4.3. Logistic Regression.
4. 4.4. Linear Discriminant Analysis
5. 4.4. Quadratic Discriminant Analysis
6. 4.5. Comparison of Classification Methods.

About Classification

- A special form of supervised learning.
- Categorical or qualititative responses: Yes/No; High/Median/Low; ...
- Main task: predict the class of the subject based on the inputs.
- This type of task is called classification.

- Widespread applications.
- This chapter will cover logistic regression; linear discriminant analysis and KNN.
- More advanced methods will be introduced in later chapters. They include:
  generalized additive model (gam); trees, random forests and boosting; and SVM; etc.

Examples

- Email spam detector
- Diagnose a person with a set of syndrome as virus carrier or non-carrier.
- Identify which gene, out of a million genes, is disease-causing or not.
- Judge if a trading activity is a fraud or not.

Examples: The default data

- Simulated data: 10000 individuals.
- Two inputs: income and balance (monthly)
- One output: Default (Yes or No).
- Judge if a trading activity is a fraud or not.

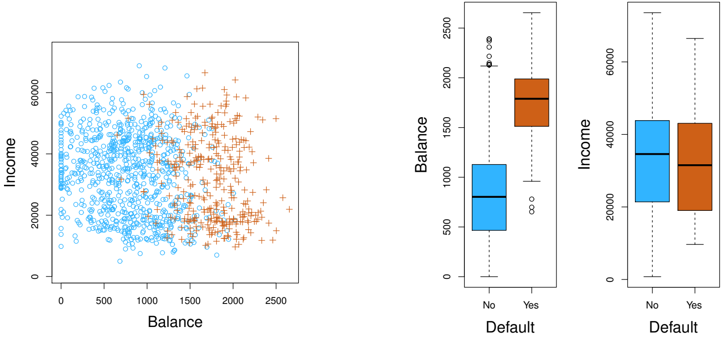


Figure: FIGURE 4.1. The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: Boxplots of income as a function of default status.

Examples: The default data

- Strong relation between balance and default.
- Weaker relation between income and default.

Why not regression

- Coding Qualitative response as numericals, such as 0, 1, 2,..., are generally inappropriate.
- classes could be Asian/European/African, Handwritten numbers (MNIST data); Identify genda (Male/female) from face image.
- In some cases, response classes are ordered, such as severe/moderate/mild; tall/medium/short.
- Coding these into 0, 1, 2 as numericals and apply regression can still be inappropriate, because it implies the differences between the adjacent classes are equal.

The special case of binary response

- Consider the output is binary: two class,
- Code the response into 0 and 1 and apply linear regression produce the same result as linear discriminant anlaysis.
- Not the case for output with more than two classes.

Example: The default data

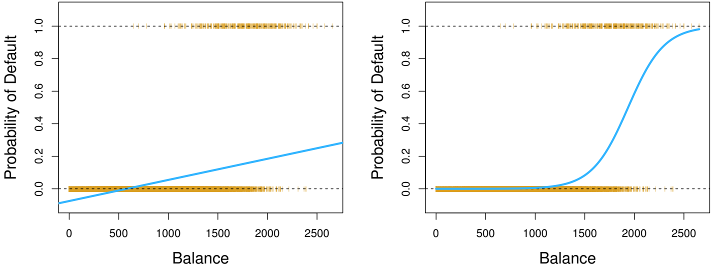


Figure: Left: linear regresion; Right: logistic regression

The setup for binary output

- The training data: $(\mathbf{x}_i, y_i)$: $i = 1, ..., n$.
- $y_i = 1$ for class 1 and $y_i = 0$ for class 0.
- $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip})$ are $p + 1$ vectors with actually p inputs.
- If instead consider linear regression model is

  $$y_i = \beta^T \mathbf{x}_i + \epsilon_i$$

  β can be estimated by the least squares, and $\hat{\beta}^T \mathbf{x}_i$ is the predictor of y_i.
- Key idea: should focus on predicting the probability of the classes.
- Using $P(y = 1|\mathbf{x}) = \beta^T \mathbf{x}$ is not appropriate.

The logistic regression model

- Assume
$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\beta^T \mathbf{x})}$$
As a result,
$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(\beta^T \mathbf{x})}$$
-
$$\log \frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})} = \beta^T \mathbf{x}$$
This is called log-odds or logit. And
$$\frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})}$$
is called odds.
- Interpretation: one unit increase in variable x_j, increases the log-odds of class 1 by β_j.

The maximum likelihood estimation

- Recall that, the likelihood is the joint probability function of joint density function of the data.
- Here, we have independent observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, each follow the (conditional) distribution

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\beta^T \mathbf{x}_i)} = 1 - P(y_i = 0|\mathbf{x}_i).$$

- So, the joint probability function is

$$\prod_{i=1,\dots,n; y_i=1} p(y_i = 1|\mathbf{x}_i) \prod_{i=1,\dots,n; y_i=0} p(y_i = 0|\mathbf{x}_i)$$

which can be conveniently written as

$$\prod_{i=1}^{n} \frac{\exp(y_i \beta^T \mathbf{x}_i)}{1 + \exp(\beta^T \mathbf{x}_i)}.$$

The likelihood and log-likelihood

- The likelihood function is the same as the joint probability function, but viewed as a function of β.
- The log-likelihood function is

$$\ell = \sum_{i=1}^{n} [y_i \beta^T x_i - \log(1 + \exp(\beta^T \mathbf{x}_i))]$$

- The maximizor is denoted as $\hat{\beta}$, which is the MLE of β based on logistic model.
- Gradient and Hessian are given as

$$\bigtriangledown \ell = \sum_{i=1}^{n} [y_i - p_i]\mathbf{x}_i = \mathbf{x}^T(\mathbf{y} - \mathbf{p}),$$

$$\mathbf{H} = -\sum_{i=1}^{n} p_i[1 - p_i]\mathbf{x}_i\mathbf{x}_i^T = -\mathbf{x}^T\mathbf{W}\mathbf{x},$$

where $\mathbf{W} = \mathrm{diag}(p_1[1 - p_1], \ldots, p_n[1 - p_n])$

The Newton-Raphson Iterative Algorithm

- The Newton step is

$$\begin{aligned}
\beta^{new} &= \beta^{old} - \mathbf{H}^{-1} \bigtriangledown \ell \\
&= \beta^{old} + (\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T(\mathbf{y}-\mathbf{p}) \\
&= (\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\left(\mathbf{x}^T\mathbf{W}\mathbf{x}\beta^{old} + \mathbf{x}^T(\mathbf{y}-\mathbf{p})\right) \\
&= (\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}\mathbf{z}
\end{aligned}$$

  where $\mathbf{z} = \mathbf{x}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})$.

- Iterative reweighted least squares.

Example: The default data with single input balance.

TABLE 4.1. (from ISLR) For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance. A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

	Coefficient	Std.error	Z-statistic	p-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
Balance	0.0055	0.0002	24.9	< 0.0001

Example: the default data with single input student

TABLE 4.2. For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable student[Yes] in the table.

	Coefficient	Std.error	t-statistic	p-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
Student[Yes]	0.4049	0.1150	3.52	0.0004

The prediction

- For the model with balance as input, predict the default probability for an individual with a balance of 1000$ as

$$\begin{aligned} & \hat{P}(default = Yes|balance = 1000) \\ = \quad & \frac{1}{1 + \exp(-(-10.6513 + 0.0055 \times 1000))} = 0.00576 \end{aligned}$$

- For the model with student as input, predict the probability of a student's default probability as

$$\begin{aligned} & \hat{P}(default = Yes|student = Yes) \\ = \quad & \frac{1}{1 + \exp(-(-3.5042 + 0.4049 \times 1))} = 0.0431 \end{aligned}$$

The prediction

- For the model with student as input, predict the probability of a non-student's default probability as

$$\begin{aligned}&\hat{P}(default = Yes|student = No)\\ = \quad &\frac{1}{1+\exp(-(-3.5042+0.4049\times 0))} = 0.0292\end{aligned}$$

- Students have higher default probabilities than non-students.

Example: the default data with all three inputs: balance, income and student

TABLE 4.3. For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance, income, and student status. Student status is encoded as a dummy variable student[Yes], with a value of 1 for a student and a value of 0 for a non-student. In fitting this model, income was measured in thousands of dollars.

	Coefficient	Std.error	t-statistic	p-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
Balance	0.0057	0.0002	24.74	< 0.0001
Income	0.0030	0.0082	0.37	0.7115
Student[Yes]	-0.6468	0.2362	-2.74	0.0062

Paradox and explanation

- Table 4.2 (single input) indicates student status has higher chance of default; but Table 4.3 (multiple input)indicates student status is negatively related with chance of default.
- Without considering other factors, students have higher chance of default.
- Considering the balance and income as additional factors, students have lower chance of default compared with non-students *who hold the same balance and income.*
- Generally, students hold higher balance (borrow more) and consequently overall higher chance of default. In other words, balance and student are two positively correlated variables.

Explanation

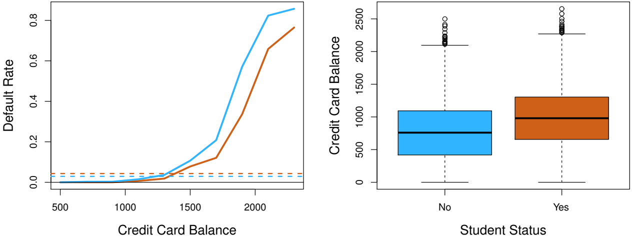
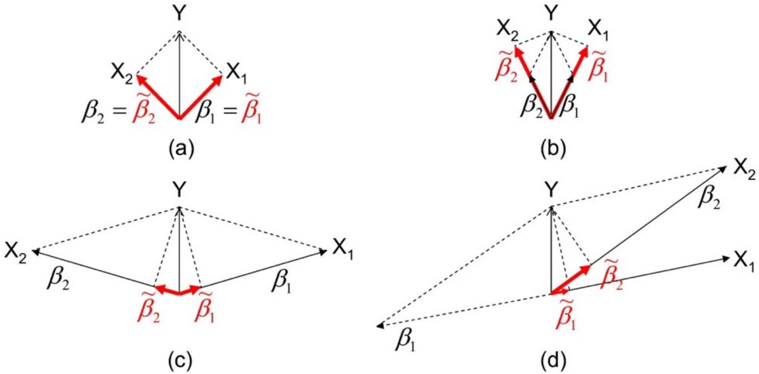


Figure 4.3. Confounding in the Default data. Left: Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of balance, while the horizontal broken lines display the overall default rates. Right: Boxplots of balance for students (orange) and non-students (blue) are shown.

A geometric explanation



Remarks

- The function $1/(1+\exp(-x))$ is called logistic sigmoid function.
- This is the cdf of the logistic distribution
- One can use other probabilistic functions to replace the sigmoid function.
- The probit model uses $\Phi(x)$, where Φ is the cdf of standard normal distribution.
- For multiple classes (more than 2 classes), the logistic regression model can be adapted by using the softmax function.

The Bayes Theorem

- Suppose there are K, denoted as $1, 2, ..., K$, for the output Y.
- X is the input of p-dimension. Both Y and X are random variables.
- Let $\pi_k = P(Y = k)$.
- Let $f_k(x) = f(x|Y = k)$ be the conditional density function of X given $Y = k$.
- Then, Bayes theorem implies

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

- General Bayes theorem :

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^K P(B|A_j)P(A_j)}$$

  for disajoint sets $A_1, ..., A_K$ whose union has probability 1.
- We classify a subject with input x into class k, if its $p_k(x)$ is the largest, for $k = 1, ..., K$.

Model assumptions of LDA

X is p-dimensional. $Y = 1, ..., K$, totally K classes. Assume, for $k = 1, ..., K$,

$$X|Y = k \sim N(\mu_k, \Sigma),$$

where μ is p-vector and Σ is p by p variance matrix. i.e.,

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\Big)$$

Note that we assumed the same Σ for all classes $k = 1, ..., K$.

Computing $p_k(x)$ for LDA

$$p_k(x) = \frac{\pi_k \exp[(-1/2)(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)]}{\sum_{l=1}^K \pi_l \exp[(-1/2)(x-\mu_l)^T \Sigma^{-1}(x-\mu_l)]}$$

Comparing $p_k(x)$ is the same as comparing the numerator. Note that the term $x^T \Sigma^{-1} x$ is common for all numerators of $p_k(x)$.
Set

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

Then a subject with input x will be classified into class k, if δ_k is the largest.
The classifer based on δ_k is the Bayesian classifier, assuming known μ_k, Σ and π_k.
A practical problem: Σ and μ_j and π_j, $j = 1, ..., K$ may be unknown.

Solution: use the sample analogue

The sample analogue of δ_k is

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

where $\hat{\mu}_k$ and $\hat{\Sigma}$ and $\hat{\pi}_k$ are sample mean, pooled sample variance, and sample proportion of class k in the data. Specifically, based on data $(x_i, y_i), i = 1, ..., n$,

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

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

where n_k is the number of subjects in class k in the data, and

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

In summary, we classify a subject with input x into class k, if $\hat{\delta}_k$ is the largest.
$\hat{\delta}_k(x)$ is called discrimination function.
$\hat{\delta}_k(x)$ is here a linear function of x. This is why we call it LDA.
The region of values of x being classified into a class has linear boundary, since $\delta_k(x) = \delta_l(x)$ defines a linear hyperplane for x.

Knowing Normal distribution

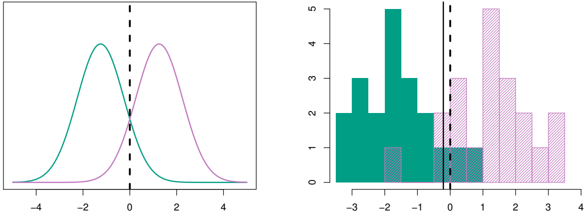


FIGURE 4.4. Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.

Knowing Normal distribution

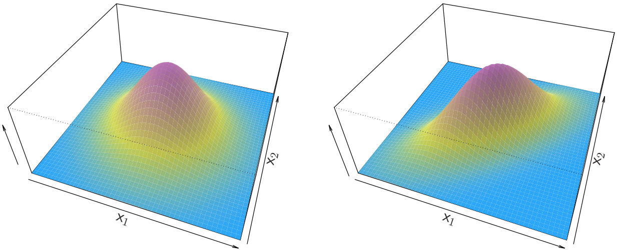
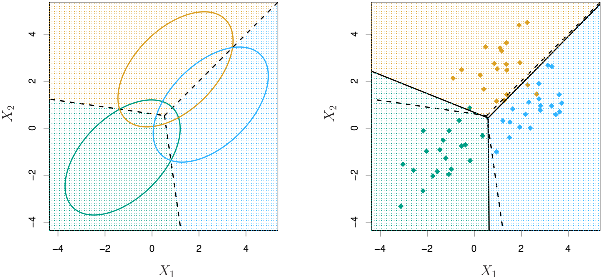


FIGURE 4.5. Two multivariate Gaussian density functions are shown, with $p = 2$. Left: The two predictors are uncorrelated. Right: The two variables have a correlation of 0.7.

The graph of LDA

FIGURE 4.6. An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with p = 2, with a class-specific mean vector and a common covariance matrix. Left: Ellipses that contain 95% of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries. Right: 20 observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.

The graph of LDA



Example: the default data

Fit the LDA model to 10,000 training samples gives training error rate of 2.75%.
That is, 275 samples are misclassifed.
But, actually only 3.33% of the training samples default, meaning that, a naive classfier that classfies every sample as non-default only has 3.33% error rate.
The detailed result is shown in the *confusion matrix.*

The Confusion matrix.

TABLE 4.4. A confusion matrix compares the LDA predictions to the true default statuses for the 10, 000 training observations in the Default data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.

		Predicted default status		
		No	Yes	Total
True	No	9644	23	9667
default status	Yes	252	81	333
	Total	9896	104	10000

Class-specific performance

- Overall error rate: $(252 + 23)/10000 = 2.75\%$.
- Error rate with default people: $252/333 = 75.7\%$
- Sensitivity: $1 - 75.7\% = 24.3\%$
- Error rate within people no-default: $23/9667 = 0.24\%$.
- Specificity: $1 - 0.24\% = 99.8\%$

A modification

- The Bayes classifier classifies a subject into class k, if the posterior probability $p_k(x)$ is the largest.
- In this case with two classes (Yes/No), i.e., $K = 2$. Bayes classifier classifies into *default* class if

$$Pr(default = Yes|X = x) > 0.5.$$

- A modification is classifies into *default* class if

$$Pr(default = Yes|X = x) > 0.2.$$

The classfication result of the modification

TABLE 4.5. A confusion matrix compares the LDA predictions to the true default statuses for the 10, 000 training observations in the Default data set, using a modified threshold value that predicts default for any individuals whose posterior default probability exceeds 20 %.

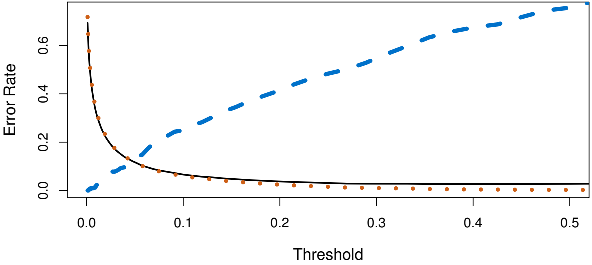
		Predicted default status		
		No	Yes	Total
True	No	9432	235	9667
default status	Yes	138	195	333
	Total	9570	430	10000

Class-specific performance

- Overall error rate: $(235 + 138)/10000 = 3.73\%$. (increased)
- Error rate with default people: $138/333 = 41.4\%$ (lower after modification)
- Sensitivity: $1 - 41.4\% = 58.6\%$
- Error rate within people no-default: $235/9667 = 2.43\%$. (increased)
- Specificity: $1 - 2.43\% = 97.57\%$
- Idenfication of defaulter (sensitivity) is more important to credit card company!
- This modification may be helpful to the campany. A tradeoff of specificity for sensitivity.

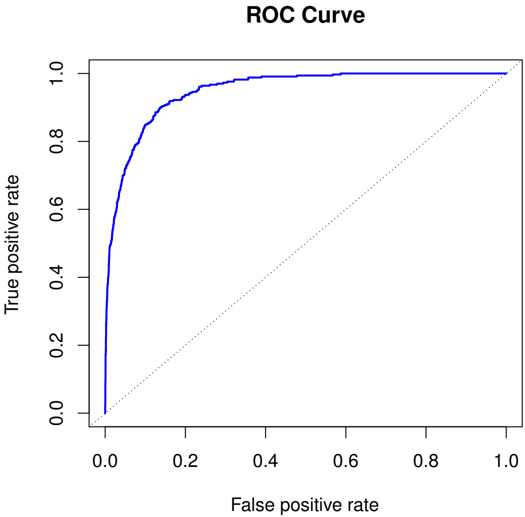
The tradeoff

FIGURE 4.7. For the Default data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are incorrectly classified, and the orange dotted line indicates the fraction of errors among the non-defaulting customers.



The ROC curve (Operation characteristic curve)

FIGURE 4.8. A ROC curve for the LDA classifier on the Default data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value. The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate. The dotted line represents the no information classifier; this is what we would expect if student status and credit card balance are not associated with probability of default. Predicted class


ROC Curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

General confusion matrix

TABLE 4.6. Possible results when applying a classifier or diagnostic test to a population.

		Predicted class		
		- or Null	+ or Non-null	Total
True status	- or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Clarifying the terminology

TABLE 4.7. Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1 - specificity
True Pos. rate	TP/P	1- Type II error, power, sensitivity, recall
Pos. Pred.value	TP/P*	Precision, 1- false discovery rate
Neg.Pred.value	TN/N*	

The model assumption of QDA

- Recall that LDA assume the class-specific normal distribution with class means μ_k and *same* class variance Σ.
- The QDA assume the class-specific normal distribution with class means μ_k and class variance Σ_k.
- The rest of the derivations follow an analogous line.

The discrimination function

- The discrimination function:

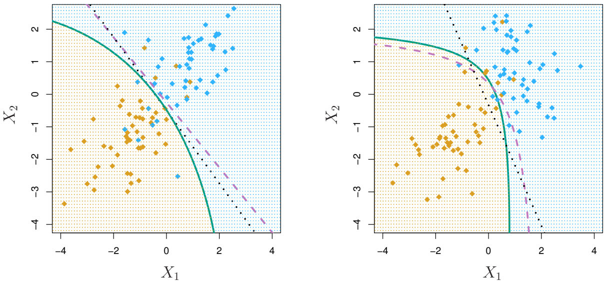
$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log(|\Sigma_k|) + \log\pi_k$$
$$= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log(|\Sigma_k|) + \log\pi_k$$

  This is a *quadratic funciton* of x.

- In actual implementation, need to use class specific sample mean and class specific sample variance to estimate μ_k and Σ_k.
- More complex than LDA, more parameters to estimate.
- **If the class variances are equal or close, LDA is better. Otherwise, QDA is better**. Is this true?

Comparing LDA and QDA

FIGURE 4.9. Left: The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with $\Sigma_1 = \Sigma_2$. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA. Right: Details are as given in the left-hand panel, except that $\Sigma_1 \neq \Sigma_2$. Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.

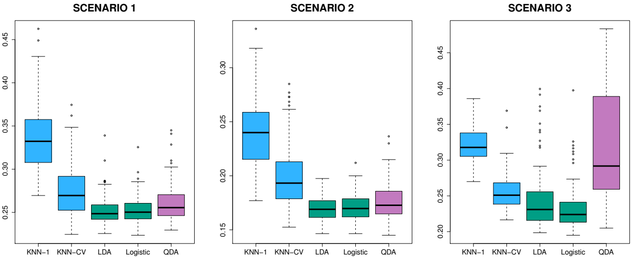


The four classification methods.

- Consider two classes $k = 1, 2$ for simplicity.
- logistic regression: $\log(p_2(x)/p_1(x)) =$ linear function of x
- LDA: $\log(p_2(x)/p_1(x)) =$ linear function of x
- The models are the same, but the methods of estimating the linear function are different, and produce different results.
- QDA: assume the log-odds are quadratic funciton of x.
- KNN.
- Comparing their performance using simulation.

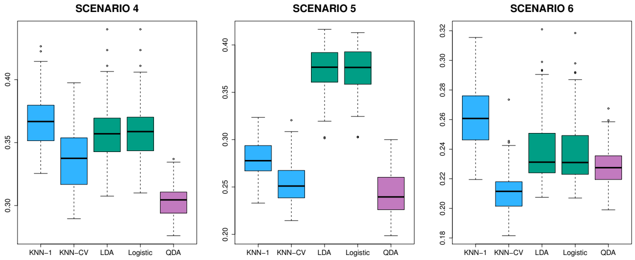
Comparions through simulation

FIGURE 4.10. Boxplots of the test error rates for each of the linear scenarios described in the main text.



Comparison through simulation

FIGURE 4.11. Boxplots of the test error rates for each of the non-linear scenarios described in the main text.



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 X1 and X2 from the *t*-distribution, with 50 observations per class. The *t*-distribution has a similar shape to distribution the normal distribution, but it has a tendency to yield more extreme pointsthat 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 considerabl 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 $X_1 \times X_2$ 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 nonlinear relationship, a non-parametric method such as KNN can still give poor results if the level of smoothness is not chosen correctly.