STA 5820
Chapter 4
Classification

Kazuhiko Shinki

Wayne State University

Overview:

- Overview
- Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- K-nearest neighbors (in Lab only)

Overview:

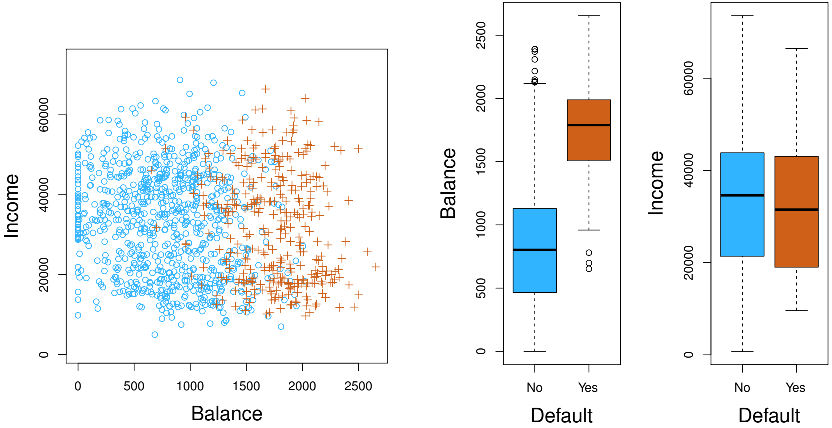
Examples of classification

- Categorize hand-written digits into classes of $\mathbf{0, 1, \cdots, 9}$.
- Categorize 150 flowers into 3 species based on measurements such as flower's diameter.
- Judge whether or not patients are malignant based on several medical measurements.

Overview:

An Example: Default data set

(Left:) Red: in default; blue: not in default. (Right:) Balance is a better indicator of default.



Why not linear regression?

- Binary classification problems can be fitted as a regression model (cf. logistic regression). (e.g., 1=malignant, 0=benign.)
- It is hard to apply a regression model to a classification problem if there are 3 categories or more and there is no presumed order (e.g., how to quantify blood type A, B, O and AB?).
  - A regression analysis is possibly fitted for each pair of categories.
- Even if there is an order for categories, it is not easy to see how to quantify the result. (e.g., Suppose that rating on movies has 1-5 scale. Distance between 1 and 2 are the same as the distance between 2 and 3?)

Logistic regression

Suppose that the response variable Y is binary ($Y = 0$ or 1), and X is a predictor variable which may be quantitative or qualitative.

Further suppose that $p(X) = P(Y = 1|X)$ (the probability that $Y = 1$ given information of X) is in $(0, 1)$ (we assume that the probability is never 0 or 1).

4.3 Logistic Regression

The logistic regression formulates $\boldsymbol{p(X)}$ as follows:

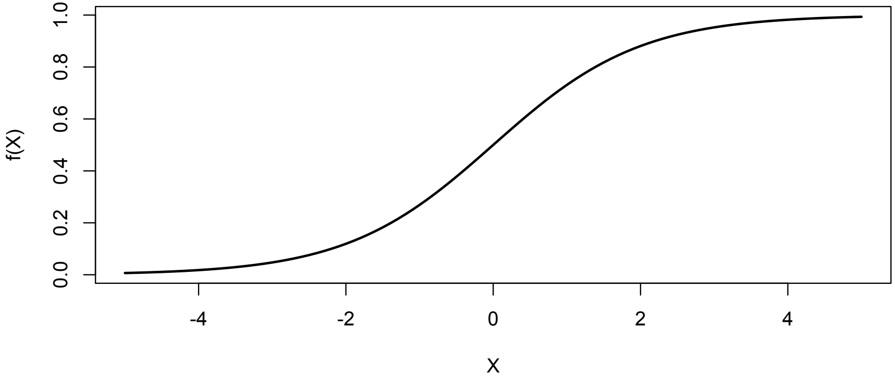
$$\boldsymbol{p(X) = f(\beta_0 + \beta_1 X)}, \quad \text{where} \quad \boldsymbol{f(a) = \frac{e^a}{1+e^a}} \quad \boldsymbol{a}, \in \mathbb{R}$$

and $\boldsymbol{f}$ is called the logistic function. Note that $\boldsymbol{f}$ is a function of $\mathbb{R} \rightarrow \boldsymbol{(0,1)}$.

This means that $\boldsymbol{p(X)}$ is explained by a linear function of $\boldsymbol{X}$ but since $\boldsymbol{p(X)}$ should be between 0 and 1, we have the function $\boldsymbol{f}$.

4.3 Logistic Regression

Figure: A graph of logistic function.



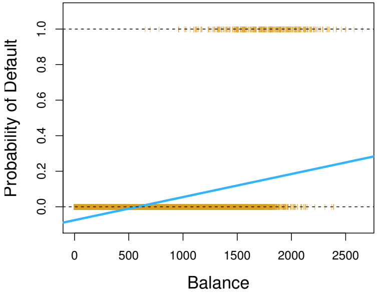
Example: 'default' data set

```
> str(Default)
'data.frame':        10000 obs. of  4 variables:
 $ default: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ student: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...
 $ balance: num  730 817 1074 529 786 ...
 $ income : num  44362 12106 31767 35704 38463 ...
> head(Default)
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559
```

Want to estimate the probability of default for each person, given their balance.

4.3 Logistic Regression

Figure 4-2: (Right:) By logistic regression model, we can estimate the default probability (blue). A larger balance implies a larger probability of default. (Left:) The model if we do not use a logistic function ***f***. The estimated probabilities may be below 0 or above 1, making poor sense.



4.3 Logistic Regression

Odds

The equation in logistic regression

$$p(X) = f(\beta_0 + \beta_1 X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

can be written as

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X},$$

which is equivalent to

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

The left-hand side is called the log-odds or logit. It is an inverse logistic function of X.

Estimating the regression coefficients

β_0 and β_1 are estimated by the maximum likelihood method.

Notion of maximum likelihood

Consider the situation to randomly pick up a die out of three ‘unfair’ dice $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ below and roll it once. Suppose you do not know which die you chose, but you can observe the result.

If the result is ‘**2**’, most likely the die is $\boldsymbol{Z}$ since $\boldsymbol{P(Z=2)}$ is larger than $\boldsymbol{P(X=2)}$ and $\boldsymbol{P(Y=2)}$. This estimator is called an maximum likelihood estimator (MLE). The function $\boldsymbol{L(\bullet)=P(\bullet=2)}$ is called a ***likelihood function*** , and the MLE is the maximizer of $\boldsymbol{L(\bullet)}$.

Probability Table

$\boldsymbol{k}$	1	2	3	4	5	6
$\boldsymbol{P(X=k)}$	1/6	1/6	1/6	1/6	1/6	1/6
$\boldsymbol{P(Y=k)}$	1/2	1/10	1/10	1/10	1/10	1/10
$\boldsymbol{P(Z=k)}$	3/10	1/2	1/20	1/20	1/20	1/20

Likelihood function of logistic regression

When $(x_1, y_1), \cdots, (x_n, y_n)$ ($y_i = 0$ or 1) are observed, the likelihood function of (β_0, β_1) is

$$
\begin{aligned}
l(\beta_0, \beta_1) &= \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j)) \\
&= \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \cdot \prod_{i:y_j=0} \frac{1}{1 + e^{\beta_0+\beta_1 x_j}}.
\end{aligned}
$$

The MLE $(\hat{\beta}_0, \hat{\beta}_1)$ is the pair of numbers which maximizes $l(\beta_0, \beta_1)$.

4.3 Logistic Regression

Table 4-1: The output of logistic regression for
$\textit{\textbf{default}} = \textit{\textbf{f}}(\beta_0 + \beta_1 \textit{\textbf{balance}})$.

	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

The coefficient are estimated by an iterative algorithm, due to a complex shape of the function $\boldsymbol{l(\beta_0, \beta_1)}$. Standard errors of coefficients are calculated by the score function. See the theory of estimation in Hastie et al. "Elements of Statistical Learning".

Making predictions

Suppose that β_0 and β_1 are estimated by $(x_1, y_1), \cdots, (x_n, y_n)$, and you want to predict y_0 for a new observation x_0. Then, the predicted value is calculated by

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_0}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_0}}.$$

Example: default data

In the above default data, $\hat{\beta}_0 = \mathbf{-10.65}$ and $\hat{\beta}_1 = \mathbf{0.0055}$. If you observe a new person with **$1,000** balance, then the predicted default probability is

$$\hat{p}(1000) = \frac{e^{-10.65+0.0055\cdot 1000}}{1 + e^{-10.65+0.0055\cdot 1000}} = 0.00576.$$

Qualitative predictor

When $\boldsymbol{X}$ is a qualitative predictor, still the logistic regression models work in the same way.

Example: Default probability by student status

Want to predict the default probability $p(X)$ by student status X ($X = 1$ if student, $X = 0$ if not). The estimates are as follows.

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

This means that

$$P(X=1) = \frac{e^{-3.5041+0.4049\cdot 1}}{1+e^{-3.5041+0.4049\cdot 1}} = 0.0431,$$

$$P(X=0) = \frac{e^{-3.5041+0.4049\cdot 0}}{1+e^{-3.5041+0.4049\cdot 0}} = 0.0292.$$

Multiple logistic regression

When there are multiple predictors $X = (X_1, \cdots, X_p)$, then the logistic regression is modeled as

$$p(X) = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p).$$

4.3 Logistic Regression

Example: dafault probability by balance, income and student status

Consider a multiple logistic regression model for default probability $p(X)$ by

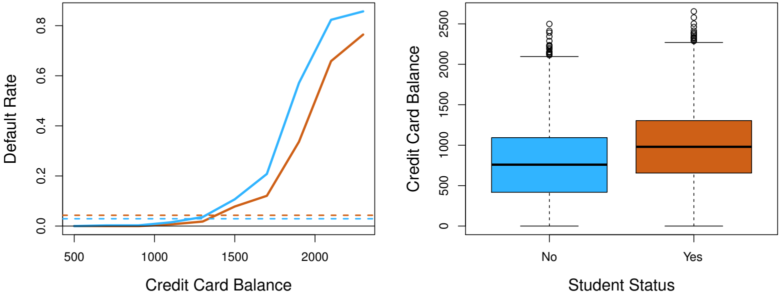
$$p(X) = f(\beta_0 + \beta_1 balance + \beta_2 income + \beta_3 student)$$

The estimated result is as follows.

	Coefficient	Std. error	Z-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

4.3 Logistic Regression

Figure 4-3: The fitted curve for **(*balance*, *default probability*)** is different depending on whether ***student*** $=$ **1** or **0**. (It is unclear what value of is `income` used to estimate the curves. Probably the mean income is used.)



Diagnosis

It is NOT meaningful to consider residual plots for logistic regression. As you can see in Figure 4-2, the pattern of a residual plot is entirely determined by the shape of the logistic curve.

This means that appropriateness of the logistic function f is largely ignored. Use of logistic function is motivated by fast estimation of parameters. In fact, the logistic function makes the likelihood function $l(\beta_0, \beta_1)$ concave, making estimation easy (see Hastie et al. "Elements of Statistical Learning").

Multinomial logistic regression: logistic regression for > 2 categories

Suppose there are K classes $1, \cdots, K$ for a response variable Y. Then, the multinomial logistic regression formulates the relationship between $P(Y=1|X), \cdots, P(Y=K|X)$ as

$$
\begin{aligned}
\log \frac{P(Y=1)}{P(Y=K)} &= \beta_{0,1} + \beta_{1,1}X_1 + \cdots + \beta_{p,1}X_p \\
&\vdots \\
\log \frac{P(Y=K-1)}{P(Y=K)} &= \beta_{0,K-1} + \beta_{1,K-1}X_1 + \cdots + \beta_{p,K-1}X_p
\end{aligned}
$$

4.3 Logistic Regression

Recall that in a logistic regression model for binary Y, the log odds is defined as

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X,$$

so the mulinomial logistic regression is a natural extension of logistic regression with $K = 2$.

Note that multinomial logistic regression with the fact that $P(Y = 1|X) + \cdots_{+} P(Y = K|X) = 1$ identifies $P(Y = 1|X), \cdots, P(Y = K|X)$.

The `multinom` function in the `nnet` package in R can estimate multinomial logistic models.

Orderded logistic regression

If a response variable Y have $K > 2$ categories which are ordered, the ordered logistic regression can fit the data. It does not give $P(Y = i|X)$ ($i = 1, \cdots, K$) anymore, but can project the corresponding class $Y = i$ conditional on X.

The `polr` function in `MASS` package in R can fit the model.

4.3 Logistic Regression

Alternative choices for logistic function

The logistic function f is popular because it is easy to use analytically. For example, it is easy to show $l(\beta_0, \beta_1)$ is a concave function, and the inverse logistic function f^{-1} has an analytical expression as seen above.

Another popular choice of function instead of f is a cumulative function of a standard normal distribution. That is,

$$\Phi(X) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$

The regression model is called Probit regression. It is a famous fact that Φ^{-1} does not have an analytical form.

Why LDA?

- LDA is more stable than logistic regression. The logistic regression is unstable when classes are well-separated in the predictor space.
- LDA is more natural when there are more than 2 classes.

4.4 Linear Discriminant Analysis

Idea of LDA

- Estimate the distribution of predictor X for each class $k = 1, \cdots, K$ (as a normal distribution).
- Use Bayes Theorem to calculate $P(Y = k|X = x)$ for $k = 1, \cdots, K$.

Example

Want to classify animals into horses, giraffes and deer by weight, height and neck length.

- Approximate the joint distribution of $\boldsymbol{X = (weight, height, necklength)}$ for each species.
- Given measurements of $\boldsymbol{(weight, height, necklength)}$, calculate the probability that the animal is a horse, a giraffe, or a deer by Bayes Theorem.

Bayes Theorem

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

(Proof:) Immediate by the definition of conditional probability:
$P(A|B) = P(A \text{ and } B)/P(B)$. $\square$

Calculating the probability for each class

Suppose $f_k(x) = P(X = x|Y = k)$ is the distribution of the predictor X given a class k. Further, let π_k denote the unconditional probability to observe class k (e.g., the proportion of the number of horses to the number of all three animals). Then,

$$\begin{aligned} Pr(Y = k|X = x) &= \frac{P(X = x, Y = k)}{P(X = x)} \\ &= \frac{P(Y = k)P(X = x|Y = k)}{\sum_{l=1}^{K} P(Y = l)P(X = x|Y = l)} \\ &= \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \end{aligned}$$

4.4.2 LInear Discriminant Analysis for $p = 1$

Suppose that X is one dimensional (e.g., only weigtht is available for animals), and suppose that f_k is Guassian (i.e., normal). Then,

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

LDA assumes σ_k does not depend on k, namely, $\sigma_1 = \cdots = \sigma_K = \sigma$.

By Bayes Theorem, it follows that

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

Bayes classifier (cf. Chapter 2.2.3) assigns the class k to x so that $p_k(x)$ is the largest among $p_1(x), \cdots, p_K(x)$.

4.4.2 LInear Discriminant Analysis for $p = 1$

Where is the boundary between two classes?

Suppose that $K = 2$ and μ_k's ($k = 1, 2$) are estimated. Where is the boundary between classes $k = 1$ and $k = 2$? As imagined, the boundary is the midpoint $(\mu_1 + \mu_2)/2$.

To see this, take the logarithm of $f_k(x)$:

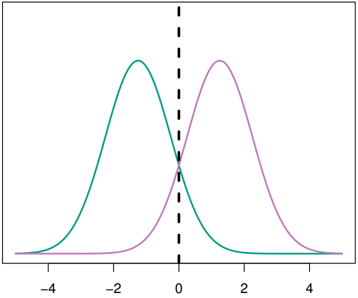
$$\delta_k(x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \mu_k)^2$$

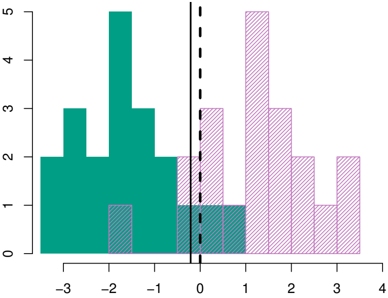
and solve $\delta_1(x) = \delta_2(x)$. Then,

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

4.4.2 LInear Discriminant Analysis for $p = 1$

Figure 4.4: Illustration of LDA boundary between the classes 1 and 2.





4.4.2 LInear Discriminant Analysis for $p = 1$

How to estimate μ_k's and σ?

μ_k's and σ are estimated by as follows.

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

$\hat{\mu}_k$ is the class mean, and $\hat{\sigma}$ is a so-called pooled standard deviation. These are unbiased estimators of the parameters.

4.4.3 LInear Discriminant Analysis for $p > 1$

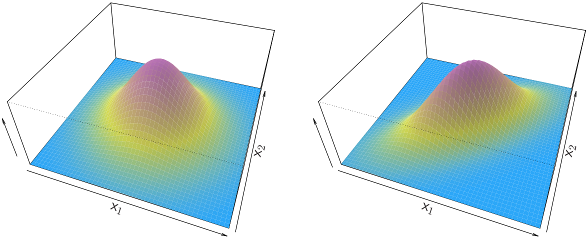
When $\boldsymbol{X}$ is $\boldsymbol{p}$-dimensional, LDA is done in the same way as 1-dimensional case but with a multivariate normal (Gassian) distribution.

We say "bivariate" normal when $\boldsymbol{p} = \mathbf{2}$.

4.4.3 LInear Discriminant Analysis for $p > 1$

Multivariate normal (Gaussian) distribution

Figure 4-5: Bivariate normal distribtuions. (Left:) Uncorrelated. $\mathbf{\Sigma}$ is diagonal. (Right:) Correlated. $\mathbf{\Sigma}$ is not diagonal.



4.4.3 LInear Discriminant Analysis for $p > 1$

Let $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{p}}$ be a column vector of predictors, $\boldsymbol{\mu} \in \mathbb{R}^{\boldsymbol{p}}$ be the population mean vector, $\boldsymbol{\Sigma} \in \mathbb{R}^{\boldsymbol{p}\times\boldsymbol{p}}$ be the population variance-covariance matrix (a positive semi-definite matrix). Then, a multivariate normal density is defined by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

where $|\bullet|$ represents the determinant of a matrix $\bullet$, and $\boldsymbol{T}$ represent transpose.

4.4.3 LInear Discriminant Analysis for $p > 1$

The LDA assumes the density function $f_k(x)$ of x given class k is

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

that is, the mean vector μ_k depends on the class, but the variance Σ does not depend on k.

Similarly to the one-dimensional case, the log of $f_k(x)$ is given by

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

where the ***constant*** does not depend on x and μ_k.

4.4.3 LInear Discriminant Analysis for $p > 1$

Decision boundary?

$\delta_k(x) = \delta_l(x)$ gives the decision boundary between the classes k and l. That is,

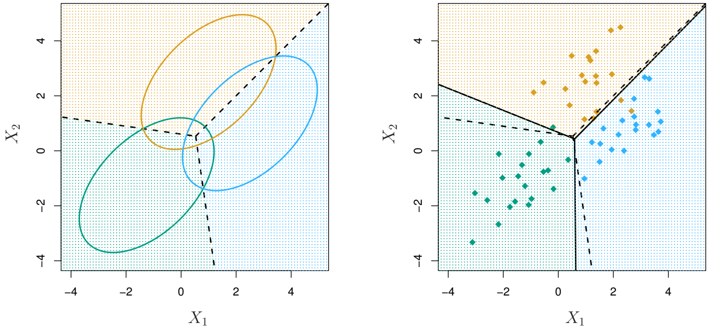
$$x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k = x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l$$

Note that the quadratic terms of x were cancelled. This boundary is a linear function of X, and hence the decision boundary is linear.

Note that the decision boundary is determined for each pair of (k, l).

4.4.3 LInear Discriminant Analysis for $p > 1$

Fig 4-6: Decision boundary given for three classes for $\boldsymbol{p = 2}$ by simulation. (Left:) The true distribution of three classes. The ellipses include 95% of observations in each class. Dashed lines represents the true optimal boundary. (Right:) Simulated data with estimated boundaries (solid black) with the true optimal boundary (dashed black).



4.4.3.b Evaluating classification performance

Consider a general problem to measure the classification performance.

Suppose that there are 10,000 people for two classes: default and no default, and the estimated classification rule mis-classified only 275 of them correctly. The error rate is 2.75%. Is it low?

4.4.3.b Evaluating classification performance

There are two possible issues.

- The training error is small, but the test error is much larger. This is especially true when the classification model is complex (overfitting).
- If a majority of observations are in one of the class, the error rate should be low. For example, in the following case, the error rate is 3.33% even if we classify all observations to the "no default" class.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

4.4.3.c Evaluating binary classification performance

We will study a few measures to evaluate binary decision rules.

First, the confusion table (or, contingency table, table for counts) is summarized below. Note that each row has a total probability of one.
Table 4-6:

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

Sensitivity & Specificity, Type I & II errors

- Sensitivity is the true positive rate, that is, TP/P.
- Specificity is the true negative rate, that is, TN/N.
- Type I error is the false positive rate, that is, FP/N or 1 - specificity.
- Type II error is the false negative rate, that is, FN/P or 1 - sensitivity.

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

4.4.3.c Evaluating binary classification performance

Example: Sensitivity & Specificity, Type I & II errors

Define default as positive. Then,

$$\begin{aligned}
\textit{sensitivity} &= 81/333 \\
\textit{specificity} &= 9644/9667 \\
\textit{Type I error} &= 23/9667 = 1 - \textit{specificity} \\
\textit{Type II error} &= 252/333 = 1 - \textit{sensitivity}
\end{aligned}$$

	Predicated as No Default	Predicted as Default	Total
No Default	9,644	23	9,667
Default	252	81	333
Total	9,896	104	10,000

4.4.3.c Evaluating binary classification performance

ROC curve

The table above classifies a person as 'Default' if

$$Pr(default = YES|X = x) > 0.5$$

If we want to have a higher sensitivity at the cost of lower sensitivity, one can classify a person as 'Default' if

$$Pr(default = YES|X = x) > 0.2$$

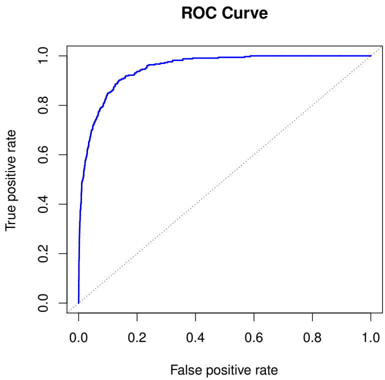
4.4.3.c Evaluating binary classification performance

Then the table becomes:

	Predicated as No Default	Predicted as Default	Total
No Default	9,432	235	9,667
Default	138	195	333
Total	9,570	430	10,000

4.4.3.c Evaluating binary classification performance

When we change the threshold probability (0.5 and 0.2 in the previous slide) little by little from one to zero, we can make a plot of all possible combinations of $\boldsymbol{(1 - \textit{specificity}, \textit{sensitivity})}$. This is called an ROC (Receiver Operating characteristics) curve.



4.4.3.c Evaluating binary classification performance

$(\textbf{\textit{False Positive rate}}, \textbf{\textit{True Positive rate}}) = (\mathbf{0}, \mathbf{1})$ is ideal, but there is a trade-off between these two.

The AUC (area under the ROC curve) is a good measure to compare different classification models.

The AUC is between 0.5 and 1, and a larger AUC is better.

The AUC is 0.95 in the above figure.

QDA

The quadratic discriminant analysis (QDA) is the same as LDA except for the fact that QDA allows different classes to have different covariance matrics $\mathbf{\Sigma}_k$.

The log of $f_k(x)$ becomes

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

For each given x, the class of k such that $\delta_k(x)$ is the largest will be assigned.

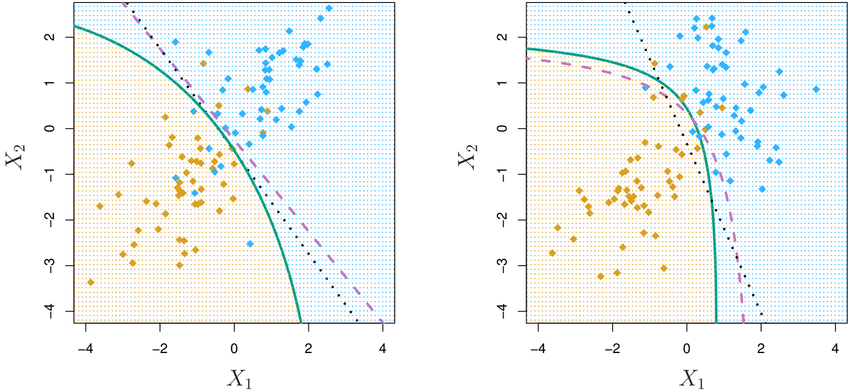
4.4.4 Quadratic Discriminant Analysis

LDA vs QDA

- Boundaries between classes are linear for LDA, hyperbolic for QDA.
- There is bias-variance trade-off between LDA and QDA.
  - QDA is more flexible than LDA.
  - QDA is more complex than LDA. Each $\boldsymbol{\Sigma}_k$ needs $\boldsymbol{p(p+1)/2}$ parameters.

4.4.4 Quadratic Discriminant Analysis

Figure 4-9: LDA vs QDA. Purple dashed = truth curve behind simulation. Black dashed = LDA. Green solid = QDA.



4.5 Comparison

Want to compare characteristics and performance of classification models (with binary classification examples).

- Logistic regression (LR)
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- K-Nearest Neighborhood (KNN)

Characteristics

- Both LR and LDA have a linear decision boundary, but estimation methods are different.
  - LR is based on a logistic function.
  - LDA is based on maximum likelihood of Gaussian densities.
- Complexity:
  - KNN (small K) > KNN (large K) > QDA > (LDA and LR).
- Forecasting performance:
  - Simple (e.g., linear) boundary: (best) LDA and LR > QDA > KNN (worst)
  - Complicated boundary: (best) KNN > QDA > LDA and LR (worst)

Model prediction performance by simulation

We will see prediction performance (i.e., classification error rate) of classification models for binary classification problems under 6 different scenarios by simulations.

The predictor is 2-dimensional continuous variable (X_1, X_2), and the response is binary.

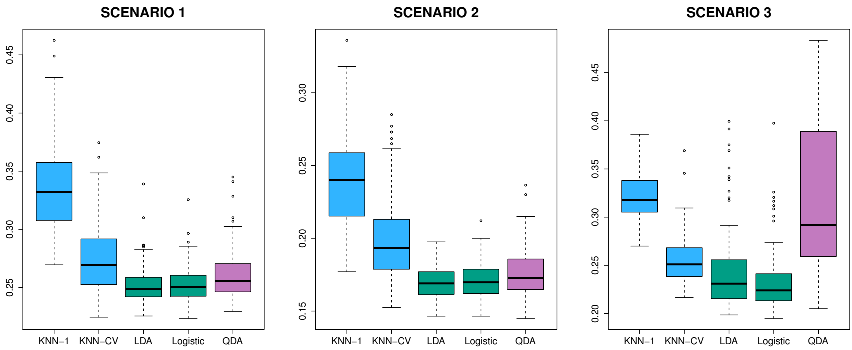
4.5 Comparison

Scenarios The true data generation process is set as follows:

- Scenario 1: Gaussian densities with no correlation b/w $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ (LDA).
- Scenario 2: Gaussian densities with positive correlation b/w $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ (LDA).
- Scenario 3: t-densities with no correlation b/w $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ (similar to LR).
- Scenario 4: Gaussian densities with different correlation b/w $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ (QDA).
- Scenario 5: $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are uncorrelated, but the responses are determined by a logistic regression with $\boldsymbol{(X_1^2, X_2^2, X_1 X_2)}$ as predictors (similar to QDA).
- Scenario 6: By a complicated rule (KNN expected to work better).

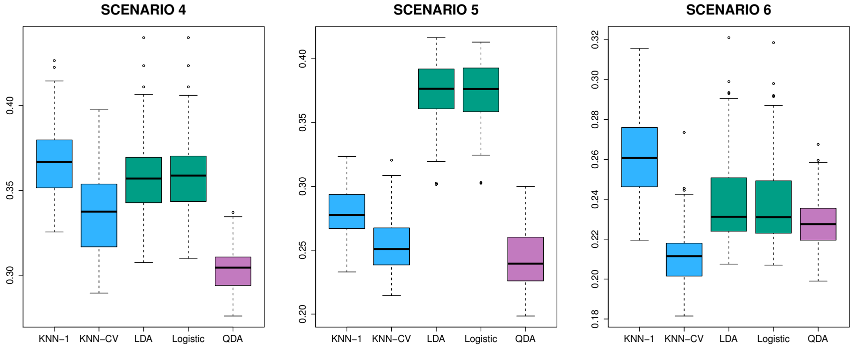
4.5 Comparison

Performance comparison: Scenarios 1-3



4.5 Comparison

Performance comparison: Scenarios 4-6



Memo