ISEP
Machine Learning
September 16th 2024

CLASSIFICATION : DISCRIMINANT ANALYSIS STEP BY STEP

62705 GUO Xiaofan

1. Exercises on Discriminant Analysis

1.1-1 Probabilities πk:

$$\pi_G = \frac{Number\ of\ G}{Total} = \frac{10}{20} = 0.5$$

$$\pi_R = \frac{Number\ of\ R}{Total} = \frac{10}{20} = 0.5$$

Mean: $\mu_G = \frac{(-1.6)+(-0.8)+(-1.8)+0.6+(-0.7)+(-1.8)+(-0.5)+(-0.3)+(-0.4)+(-1.3)}{10} = \frac{-8.6}{10} = -0.86$

$$\mu_R = \frac{2.5 + 1.4 + 0.4 + (-1.2) + 2.1 + 1.0 + 1.0 + 1.9 + 1.8 + 1.6}{10} = \frac{12.5}{10} = 1.25$$

Var: $\sigma^2 = \frac{1}{20-2}\sum_{k=1}^{2}\sum_{i:y_i=k}(x_i - \mu_k)^2 = \frac{1}{20-2}\left(\sum_{i:y_i=G}(x_i - \mu_G)^2 + \sum_{i:y_i=R}(x_i - \mu_R)^2\right)$

$$= \frac{1}{18}[(-1.6 + 0.86)^2 + (-1.3 + 0.86)\text{^}2 + (2.5 - 1.25)^2 + \text{^} + (1.6 - 1.25)^2]$$

$$= \frac{1}{18}(5.7 + 11.7) = 0.9667$$

$$\sigma = \sqrt{0.9667} = 0.9832$$

The density of X: $f_G(x) = \frac{1}{\sqrt{2\pi}*\sigma_G} e^{-\frac{1}{2}(\frac{x-\mu_G}{\sigma_G})^2} = \frac{1}{\sqrt{2\pi}*0.9832} e^{-\frac{1}{2}(\frac{x+0.86}{0.9832})^2}$

$$f_R(x) = \frac{1}{\sqrt{2\pi} * \sigma_R} e^{-\frac{1}{2}(\frac{x-\mu_R}{\sigma_R})^2} = \frac{1}{\sqrt{2\pi} * 0.9832} e^{-\frac{1}{2}(\frac{x-1.23}{0.9832})^2}$$

1.1-2 pˆk(x)

$$P(Y = G|X = x) = p_G(X = x) = \frac{0.5 * f_G(x)}{0.5 * f_G(x) + 0.5 f_R(x)} = \frac{f_G(x)}{f_G(x) + f_R(x)}$$

$$P(Y = R|X = x) = p_R(X = x) = \frac{0.5 * f_R(x)}{0.5 * f_R(x) + 0.5 f_G(x)} = \frac{f_R(x)}{f_R(x) + f_G(x)}$$

Give x=0.5:

$$f_G(0.5) = \frac{1}{\sqrt{2\pi} * 0.9832} e^{-\frac{1}{2}(\frac{0.5+0.86}{0.9832})^2} = 0.156$$

$$f_R(0.5) = \frac{1}{\sqrt{2\pi} * 0.9832} e^{-\frac{1}{2}(\frac{0.5-1.23}{0.9832})^2} = 0.304$$

Then:

$$P(Y = G|X = 0.5) = \frac{0.156}{0.156 + 0.304} = 0.339$$

$$P(Y = R|X = 0.5) = \frac{0.304}{0.304 + 0.156} = 0.661$$

1.1-3 The function lda() in R:

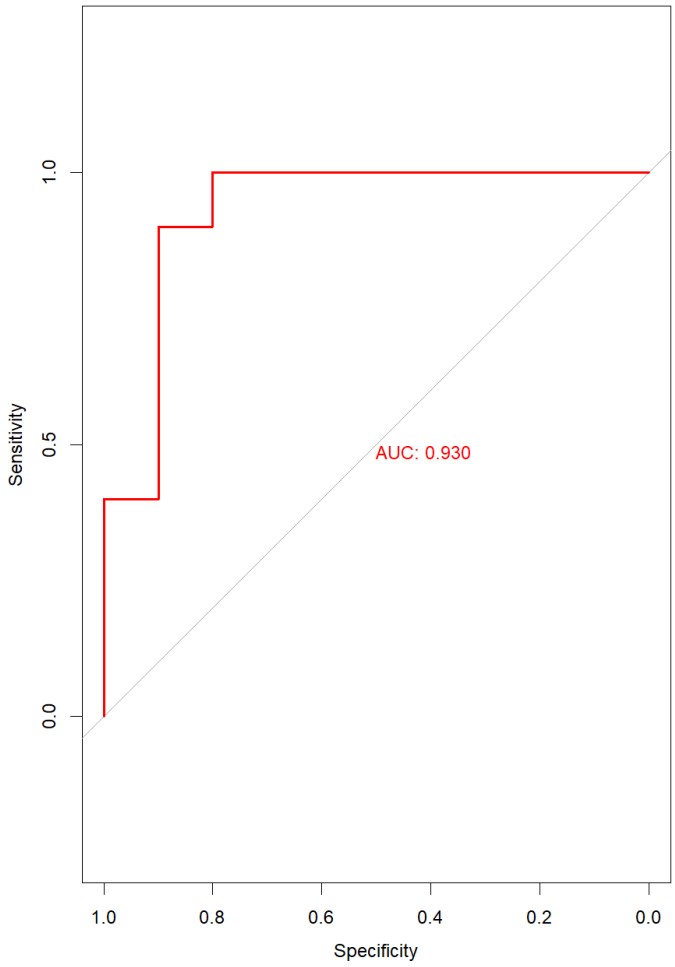
```
> install.packages("MASS")
WARNING: Rtools is required to build R packages but is not currently insta
led. Please download and install the appropriate version of Rtools before
roceeding:

https://cran.rstudio.com/bin/windows/Rtools/
将程序包安装入'C:/Users/16273/AppData/Local/R/win-library/4.4'
(因为'lib'没有被指定)
试开URL'https://cran.rstudio.com/bin/windows/contrib/4.4/MASS_7.3-61.zip'
Content type 'application/zip' length 1166734 bytes (1.1 MB)
downloaded 1.1 MB

程序包'MASS'打开成功，MD5和检查也通过

下载的二进制程序包在
        C:\Users\16273\AppData\Local\Temp\RtmpKmQ84x\downloaded_packages里
> library(MASS)
> X <- c(-1.6, -0.8, -1.8, 0.6, -0.7, -1.8, -0.5, -0.3, -0.4, -1.3, 2.5, 
4, 0.4, -1.2, 2.1, 1.0, 1.0, 1.9, 1.8, 1.6)
> Y <- factor(c(rep("G", 10), rep("R", 10)))
> lda.fit <- lda(Y ~ X)
> lda.pred <- predict(lda.fit)
> lda.pred$posterior
            G           R
1  0.98842647 0.011573528
2  0.92167086 0.078329143
3  0.99291683 0.007083171
4  0.26826346 0.731736540
5  0.90181029 0.098189708
6  0.99291683 0.007083171
7  0.84838259 0.151617406
8  0.77319490 0.226805102
9  0.81369580 0.186304202
10 0.97596925 0.024030750
11 0.00329845 0.996701550
12 0.04808176 0.951918245
13 0.37568134 0.624318661
14 0.96941939 0.030580611
15 0.00883696 0.991163040
16 0.11978016 0.880219844
17 0.11978016 0.880219844
18 0.01442298 0.985577023
19 0.01840351 0.981596493
20 0.02985458 0.970145420
> print(lda.pred$posterior[, 1])
         1          2          3          4          5          6
0.98842647 0.92167086 0.99291683 0.26826346 0.90181029 0.99291683
         7          8          9         10         11         12
0.84838259 0.77319490 0.81369580 0.97596925 0.00329845 0.04808176
        13         14         15         16         17         18
0.37568134 0.96941939 0.00883696 0.11978016 0.11978016 0.01442298
        19         20
0.01840351 0.02985458
> print(lda.pred$posterior[, 2])
         1          2          3          4          5          6
0.011573528 0.078329143 0.007083171 0.731736540 0.098189708 0.007083171
         7          8          9         10         11         12
0.151617406 0.226805102 0.186304202 0.024030750 0.996701550 0.951918245
        13         14         15         16         17         18
0.624318661 0.030580611 0.991163040 0.880219844 0.880219844 0.985577023
        19         20
0.981596493 0.970145420
> 
```

For x=0.5(the 4th number):

	R	CALCULATE
$P(Y = G\|X = 0.5)$	0.26826346	0.339
$P(Y = R\|X = 0.5)$	0.731736540	0.661

This may be due to rounding during calculation.

1.2 Measuring the accuracy with a ROC curve



1) The vertical axis represents Sensitivity: shows the classifier's ability to correctly identify positive cases.
2) The horizontal axis represents Specificity (False Positive Rate): shows the proportion of negative cases incorrectly classified as positive.
3) The shape of the ROC curve indicates the trade-off between sensitivity and specificity at different thresholds.
4) The AUC is 0.930, which means the classifier has excellent overall performance.