M072040019 梅瀚中

7. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80 % of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. *You will need to use Bayes' theorem.*

$$P_k(x) = \frac{0.8 \times \frac{1}{\sqrt{2\pi}6} e^{\frac{-(4-10)^2}{72}}}{0.8 \times \frac{1}{\sqrt{2\pi}6} e^{\frac{-(4-10)^2}{72}} + 0.2 \times \frac{1}{\sqrt{2\pi}6} e^{\frac{-(4-0)^2}{72}}}$$

$$= \frac{0.8\, e^{\frac{-36}{72}}}{0.8\, e^{\frac{-1}{2}} + 0.2\, e^{\frac{-16}{72}}}$$

$$= 0.7518$$

```
> (0.8*exp(-1/2))/(0.8*exp(-1/2)+0.2*exp(-16/72))
[1] 0.7518525
```

8. Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20 % on the training data and 30 % on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training data sets) of 18 %. Based on these results, which method should we prefer to use for classification of new observations? Why?

使用羅吉斯做預測比較好,因為 knn 有平均 18%的預測錯誤率,假使訓練集的錯誤率為 0%,結果測試集錯誤率達 36%,比羅吉斯的 30%在測試的錯誤率還高。

9. This problem has to do with *odds*.

   (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

   (b) Suppose that an individual has a 16 % chance of defaulting on her credit card payment. What are the odds that she will default?

(a)

$$\frac{p(x)}{1-p(x)} = 0.37$$

所以,

$$p(X) = \frac{0.37}{1+0.37} = 0.27$$

(b)

$$\frac{p(x)}{1-p(x)} = \frac{0.16}{1-0.16} = 0.1904672$$

12. This problem involves writing functions.

    (a) Write a function, `Power()`, that prints out the result of raising 2 to the 3rd power. In other words, your function should compute 2^3 and print out the results.

    Hint: Recall that `x^a` *raises* `x` *to the power* `a`. *Use the* `print()` *function to output the result.*

    (b) Create a new function, `Power2()`, that allows you to pass *any* two numbers, `x` and `a`, and prints out the value of `x^a`. You can do this by beginning your function with the line

    ```
    > Power2=function(x,a){
    ```

    You should be able to call your function by entering, for instance,

    ```
    > Power2(3,8)
    ```

    on the command line. This should output the value of 3^8, namely, $6,561$.

    (c) Using the `Power2()` function that you just wrote, compute 10^3, 8^{17}, and 131^3.

    (d) Now create a new function, `Power3()`, that actually *returns* the result `x^a` as an `R` object, rather than simply printing it to the screen. That is, if you store the value `x^a` in an object called `result` within your function, then you can simply `return()` this result, using the following line:

```
return(result)
```

The line above should be the last line in your function, before the } symbol.

(e) Now using the `Power3()` function, create a plot of $f(x) = x^2$. The x-axis should display a range of integers from 1 to 10, and the y-axis should display x^2. Label the axes appropriately, and use an appropriate title for the figure. Consider displaying either the x-axis, the y-axis, or both on the log-scale. You can do this by using `log=''x''`, `log=''y''`, or `log=''xy''` as arguments to the `plot()` function.

(f) Create a function, `PlotPower()`, that allows you to create a plot of `x` against `x^a` for a fixed `a` and for a range of values of `x`. For instance, if you call

```
> PlotPower(1:10,3)
```

then a plot should be created with an x-axis taking on values $1, 2, \ldots, 10$, and a y-axis taking on values $1^3, 2^3, \ldots, 10^3$.

(a)

```
> PlotPower(1:10, 3)
> Power <- function(){
+   2^3
+ }
> Power()
[1] 8
```

(b)

```
> Power2=function(x,a){
+   x^a
+ }
> Power2(3,8)
[1] 6561
```

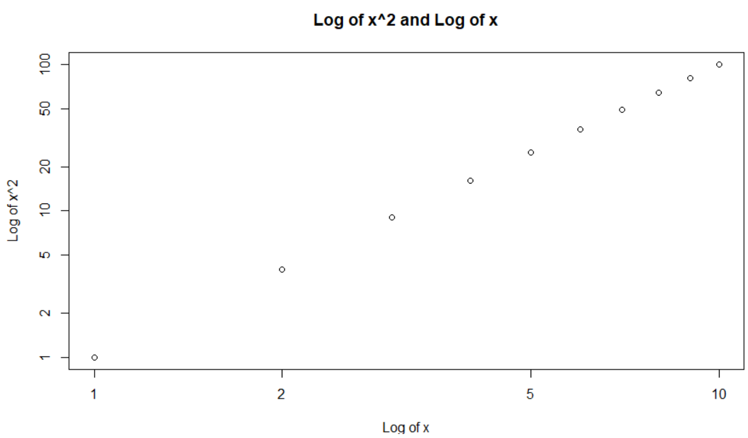
(C)

```
> Power2(10,3)
[1] 1000
> Power2(8,17)
[1] 2.2518e+15
> Power2(131,3)
[1] 2248091
```

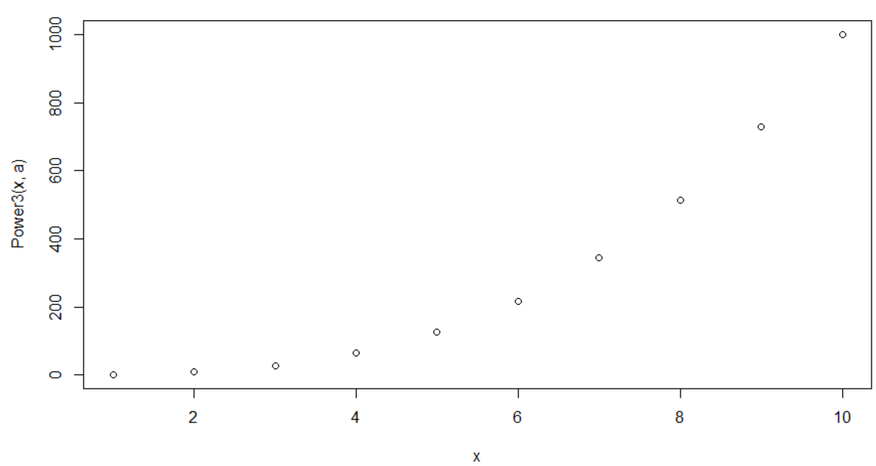
(d)

```
> Power3 <- function(x , a) {
+   result <- x^a
+   return(result)
+ }
```

(e)



(f)



13. Using the Boston data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, and KNN models using various subsets of the predictors. Describe your findings.

```
> #13
> library(MASS)
> data("Boston")
> crim01 <- rep(0, length(Boston$crim))
> crim01[Boston$crim > median(Boston$crim)] <- 1
> Boston <- data.frame(Boston, crim01)
>
>
> #logistic
> fit.glm1 <- glm(crim01 ~ . - crim01 - crim, data = Boston, family = bin
omial)
> probs <- predict(fit.glm1, Boston.test, type = "response")
> pred.glm <- rep(0, length(probs))
> pred.glm[probs > 0.5] <- 1
> table(pred.glm, crim01.test)
        crim01.test
pred.glm  0  1
       0 67  7
       1  5 73
> #lda
> fit.lda <- lda(crim01 ~ . - crim01 - crim , data = Boston)
> pred.lda <- predict(fit.lda, Boston.test)
> table(pred.lda$class, crim01.test)
   crim01.test
     0  1
  0 66 18
  1  6 62
>
```