STAT5104 16/17 Second term Final Examination

Answer **ALL** Questions (Time: 2 hour). Show all the detail of your calculation. Hand in this question paper together with your answer book.

Dataset for Question 1 to 4

The following dataset are from the 1995 US News report on American colleges and universities. with the following variables:

Column	Name	Description
1	app	No. of applicants received
2	acc	No. of applicants accepted
3	enrol	No. of new students enrolled
4	ftime	No. of full-time undergraduates
5	ptime	No. of part-time undergraduate
6	instate	In-state tuition
7	outstate	Out-state tuituion
8	rbcost	Room and board cost
9	bkcost	Estimated book cost
10	phd	% of faculty with Ph.D.
11	sfratio	Student/faculty ratio
12	expend	Instructional expenditure per student
13	grad	Graduation rate
14	top10	% new students from top 10% of high school class

The dataset `d` are randomly partitioned into training dataset `d0` and testing dataset `d1`. The last column is transform into a binary variable y0 and y1 and used as a target variable as follow:

```
y0<-(d0$top10>30)+0       # create target var y0=1 if top10>30; y0=0 otherwise
```

Question 1 [20%]

A CTREE is applied with the following output:

```
ctree<-rpart(y0~.,data=d0[,c(1:13)],method="class",maxdepth=3)
print(ctree)
1) root 700 191 0 (0.72714286 0.27285714)
   2) instate< 12630 534  75 0 (0.85955056 0.14044944)
     4) phd< 79.5 425  37 0 (0.91294118 0.08705882) *
     5) phd>=79.5 109  38 0 (0.65137615 0.34862385)
      10) ptime>=111.5 96  26 0 (0.72916667 0.27083333) *
      11) ptime< 111.5 13   1 1 (0.07692308 0.92307692) *
   3) instate>=12630 166  50 1 (0.30120482 0.69879518)
     6) expend< 11967.5 80  40 0 (0.50000000 0.50000000)
      12) ptime>=163 44  14 0 (0.68181818 0.31818182) *
      13) ptime< 163 36  10 1 (0.27777778 0.72222222) *
     7) expend>=11967.5 86  10 1 (0.11627907 0.88372093) *
```

(a) Draw the classification tree and produce the classification table.

(b) Write down the rule with the highest confidence. What is the support and confidence of this rule? Is this rule useful? Explain your answer.

(c) If a record randomly selected, what is the probability that top10>30 in that record?

(d) Consider the rule R: If (instate>=12630) then (top10>30). What is the confidence and lift value of this rule?

(e) If we know that `top10` in a selected record is greater than 30, what is the probability that `instate<12630`?

Question 2 [20%]

ANN is applied to the dataset with `ptime`, `instate`, `phd`, `expend` (i.e, columns 5,6,10,12) as input with size=2 and **logistic** is used. The following are the R commands and output:

```
y0<-factor(y0)
col.nn<-ann(d0[,c(5,6,10,12)],y0,size=2,try=30)
summary(col.nn)

 b->h1 i1->h1 i2->h1 i3->h1 i4->h1
  0.30   0.10   0.24   0.56   0.33
 b->h2 i1->h2 i2->h2 i3->h2 i4->h2
 -0.59  -0.36   0.53  -0.25  -0.25
 b->o h1->o h2->o
-1.37 -0.41  1.03
```

(a) Write down exactly the system of equations of this ANN model.

(b) Suppose we have a record: (i1,i2,i3,i4)=(ptime,instate,phd,expend)=(869,7560,76,10922). What is the probability that `top10>30` in this record?

(c) If we change the size from 2 to 4, what is the number of parameters in this new ANN model? What is the potential problem with this new ANN model? Be specific.

(d) Is it possible to use top10 as the target variable instead of using y0 in ANN? Do NOT write any R codes, just explain your answer clearly.

Question 3 [20%]

A binary variable `ins` is created from `instate` using the following R command:

```
ins<-(d0$instate>5000)+0       # ins=1 if instate>5000; ins=0 otherwise
```

The following are the R commands and output from a logistic regression:

```
summary(glm(y0~ptime+phd+expend+grad+ins,data=d0,binomial))

             Estimate   Std. Error z value Pr(>|z|)
(Intercept) -9.3422     8.731e-01 -10.700  < 2e-16 ***
ptime       -0.0008     1.779e-04  -4.270 1.95e-05 ***
phd          0.0616     1.092e-02   5.637 1.73e-08 ***
expend       0.0003     5.212e-05   6.456 1.08e-10 ***
grad         0.0260     8.535e-03   3.048  0.00231 **
ins         -0.7918     3.570e-01  -2.218  0.02658 *
```

(a) Write down the logistic regression for the two group `ins=0` and `ins=1` separately.

(b) Suppose we have a record: `(ptime,phd,expend,grad,top10)=(74,76,13965,77,50)`. Give the best prediction whether instate in this record is larger than 5000 or not. Explain your answer.

(c) If we create new binary variables **`ins1<-2*ins-1`**, **`yp<-1-y0`**, and fit a logistic regression:
`summary(glm(`**`yp`**`~ptime+phd+expend+grad+`**`ins1`**`,data=d0,binomial))`

Find the max. likelihood estimate of the regression coefficients of this logistic regression.

(d) Can we use the ordinary regression model with `top10` as the dependent variable and other variables as independent variables? Explain your answer.

Question 4 [20%]

A kmeans clustering is performed on the whole dataset d using ptime, instate, phd, top10 (i.e, columns 5,6,10,14) as input with k=2. The following are the R codes and outputs from the R built-in function kmeans():

```
x<-scale.con(d[,c(5,6,10,14)])# rescale input to [0,1] and save to x
km2<-kmeans(x,2)              # kmeans with k=2
km2$centers                   # cluster center
   ptime instate    phd  top10
1 0.1286  0.2767 0.5764 0.1795
2 0.0454  0.6937 0.7731 0.3872

km2$size                      # cluster size
[1] 517 302

km2$withinss
[1] 48.9354 22.6140
```

(a) Describe briefly the characteristic of cluster 1 and cluster 2.

(b) Compute the overall mean of the dataset x.

(c) Compute the within group SS tr(SSW) and between group SS tr(SSB).

(d) Consider a record from x: (ptime,instate,phd,top10)=(0.296,0.115,0.893,0.368). Using this kmeans clustering, should we classify this record to cluster 1 or cluster 2? Explain your answer in details.

(e) Give two suggestions to improve this kmeans clustering result. Be specific.

Question 5 [20%]

Suppose a dataset has N records with two categorical variables A and B. A has I categories $\{A_1, A_2, ...A_I\}$ and B has J categories $\{B_1, B_2, ...B_J\}$. Furthermore, let n_{ij} denotes the frequency count of $A = A_i$ and $B = B_j$. Note that $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = N$.

(a) If a record is selected at random, what is $\Pr\{A = A_i\}$ in this record? What is $\Pr\{A = A_i \mid B = B_j\}$ in this record? Under what condition is that the information of $B = B_j$ is useful in predicting $A = A_i$? Be specific.

(b) Consider the rules R1: If $A = A_i$ then $B = B_j$ and R2: If $B = B_j$ then $A = A_i$. What is the confidence and the lift value of R1 and R2 respectively?

- END OF QUESTIONS -

- Please return this question paper with your answer book -