Homework 7

Prakash Paudyal

You do not need to include the above statements.

Please do the following problems from the text book ISLR. (use set.seed(702) to replicate your results).

1. Question 5.4.1 pg 197

Using basic statistical properties of the variance, as well as singlevariable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize $Var(\alpha X + (1\alpha)Y)$

To Minimize the total risk or variance, we will minimize the $Var(\alpha X+(1\alpha)Y)$where X and Y are two random variables.Here, we have following proporties of variance , those we can use to similify our variaance.

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$
$$Var(aX) = a^2 Var(X)$$
$$Cov(aX, bY) = abCov(X,Y)$$

Then, we can use this formulas to variance of two random variables

$$Var(\alpha X + (1-\alpha)Y) = Var(\alpha X) + Var((1-\alpha)Y) + 2Cov(\alpha X, (1-\alpha)Y)$$

$$= \alpha^2 Var(X) + (1-\alpha)^2 Var(Y) + 2\alpha(1-\alpha)Cov(X,Y)$$

$$f(\alpha) = \sigma_X^2 \alpha^2 + \sigma_Y^2 (1-\alpha)^2 + 2\sigma_{XY}(-\alpha^2 + \alpha)$$

To get the minimum value of variance,that is zero, we can take fist darivates of above equation with respect to α, which is critical point for value of α.

$$\frac{d}{d\alpha} f(\alpha) = 0$$
$$\frac{d}{d\alpha} f(\alpha) = 2\sigma_X^2 \alpha + 2\sigma_Y^2 (1-\alpha)(-1) + 2\sigma_{XY}(-2\alpha + 1) = 0$$
$$2\sigma_X^2 \alpha + \sigma_Y^2(\alpha - 1) + \sigma_{XY}(-2\alpha + 1) = 0$$
$$(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\alpha - \sigma_Y^2 + \sigma_{XY} = 0$$
$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Hence this is the minimum possible value of α to minimize the given variance $Var(\alpha X + (1\alpha)Y)$

2. Question 5.4.6 pg 199

6.We continue to consider the use of a logistic regression model to predict the probability of default using income and balance on the Default data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed before beginning your analysis.

default	student	balance	income
No	No	729.5265	44361.625
No	Yes	817.1804	12106.135
No	No	1073.5492	31767.139
No	No	529.2506	35704.494
No	No	785.6559	38463.496
No	Yes	919.5885	7491.559

(a)

Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.5404684	0.4347564	-26.544680	0.00e+00
income	0.0000208	0.0000050	4.174178	2.99e-05
balance	0.0056471	0.0002274	24.836280	0.00e+00

(b)

** Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model**

```
boot.fn = function(data, index)
{
fit<-glm(default ~ income + balance,data = data, family = "binomial", subset = index)
return(coef(fit))
}
#boot.fn(Default,110)
```

(c)

** Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance.**

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Default, statistic = boot.fn, R = 100)
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1* -1.154047e+01  9.699111e-02 4.101121e-01
## t2*  2.080898e-05  6.715005e-08 4.127740e-06
## t3*  5.647103e-03 -5.733883e-05 2.105660e-04
```

```
##   (Intercept)        income       balance
## -1.154047e+01  2.080898e-05  5.647103e-03
```

Here t_1=β_0 ,t_2=β_1 ,t_3=β_2 and standard error of the logistic regression coefficients for income and balance are 0.4239, 4.583 x 10^(-6) and 2.268 x 10^(-4) respectivly

(d)

Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.

The standard error obtaind by both method are close to eachother.

3. Question 5.4.9 pg 201

9. We will now consider the Boston housing data set, from the MASS library.

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

(a)

Based on this data set, provide an estimate for the population mean of medv. Call this estimate $\hat{\mu}$

```
## [1] 22.53281
```

$\hat{\mu}$=22.53281

(b)

Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result. Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

```
## [1] 0.4088611
```

$sd.\hat{e}rr$ of $\hat{\mu}$ =0.4088611

The standard error of the mean provides a rough estimate of the interval in which the population mean is likely to fall. The population mean lies in the interval m ± 2SE, where m is sample mean.

(c)

Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = medv, statistic = boot.fnn, R = 1000)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
## t1* 22.53281 0.01674209   0.4025011
```

The standard error in b)=0.4088611 and by using bootstrap sd.error=0.4025011 which almost similar value.

(d)

Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of medv. Compare it to the results obtained using t.test(Boston$medv). Hint: You can approximate a 95% confidence interval using the formula $\hat{\mu} - 2\text{SE}(\hat{\mu})$, $\hat{\mu} + 2\text{SE}(\hat{\mu})$.

	Lower 95%	Upper 95%
Bootstrap	21.72780	23.33781
t.test	21.72953	23.33608

Bootstrap estimate for 95% confidence interval is pretty much close to t.test estimate.

(e)

Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of medv in the population.

```
## [1] 21.2
```

$\hat{\mu}_{med}$= 21.2

(f)

We now would like to estimate the standard error of ˆ med. Unfortunately,there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = medv, statistic = boot.fn2, R = 1000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1*     21.2 -0.02805   0.3711245
```

Estimated median value is similar to previous one and standard error is 0.3711245 which is smaller than the mean standard error.

(g)

Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$ (You can use the quantile() function.)

```
##   10%
## 12.75
```

$\hat{\mu}_{0.1}$=12.75

(h)

Use the bootstrap to estimate the standard error of ˆ 0.1. Comment on your findings.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = medv, statistic = boot.fn3, R = 1000)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*    12.75  0.0249   0.5002952
```

The bootstrap estimate of boot quantile is very close to estimates obtaind wiht whole dataset.The standard error is 0.5002952. Median value is same as abtaind from entire dataset.

4.

Last homework you have used different classification methods to analyze the dataset you chose.

Now use i) Validation Set Approach (VSA) ii) LOOCV and 5-fold Cross Validation

```
to estimate the test error for the following models. Choose the best model based on test error.
i) Logistic Regression (or Multinomial Logistic Regression for more than two classes)
ii) KNN (choose the best of K)
iii) LDA
iv) QDA
v) MclustDA - best model chosen by BIC
vi) MclustDA with modelType="EDDA"
vii) Find a new method that we haven't covered in class that can do classification.
```

age workclass fnlwgt education educationnum maritalstatus occupation relationship race sex capitalgain capitalloss hourspweek nativecountry income

Summarize the results in a table form (See below). **Do NOT** show your summary directly from the code. Report only the important information as figures or tables. If you can't perform any of the analysis mentioned above, write the reason why. Write a discussion and draw conclusions in the context of the original problem from your analysis. (The following table could be used, other options would be the kable() command in the knitr library, or using inline code)

	Test Error		
Method	VSA	LOOCV	5-Fold CV
Logistic Reg	0.1793391	0.1873464	0.1219478
KNN	0.2428449	0.3593366	0.2909104
LDA	0.1770053	0.2168305	0.2194654
QDA	0.2154526	0.213145	0.2096682
MclustDA	0.2117676		0.2307535
MclustDA (EDDA)	0.2067314		0.2027846
SVM	0.0002456701	0.001228501	0.0002047502

Out of all the algorithms that I experimented with, the best result (i.e. least value for test error) was obtained for 5 fold cross-validation in SVM. The test error for VSA technique using SVM is very close to our best result. MclustDA and MclustDA (EEDA) did not run as I kept on getting an error saying that some of the variables in our dataset appeared to be constant within groups.

Details work of Q4

Import the data from a url

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age          : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass    : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt       : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education    : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ educationnum : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ maritalstatus: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation   : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race         : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capitalgain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capitalloss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hoursperweek : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ nativecountry: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ income       : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

There are some categorical variables where the missing levels are coded as ? and there are more than 10 levels for some categorical variables. Hence we will relevels some of catogorical variable to reduce the number of levels and replace the level ? by misslevel.

Data preprocessing (collapse the factor levels & re-coding)

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age          : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass    : Factor w/ 9 levels "misLevel","FedGov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt       : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education    : Factor w/ 8 levels "presch","primary",..: 6 6 NA 5 6 7 4 NA 7 6 ...
##  $ educationnum : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ maritalstatus: Factor w/ 4 levels "divorce","married",..: 3 2 1 2 2 2 2 2 3 2 ...
##  $ occupation   : Factor w/ 5 levels "misLevel","clerical",..: 2 NA 3 3 3 NA 3 NA 3 NA ...
```

```
##  $ relationship : Factor w/ 6 levels "husband","wife",..: 3 1 3 1 2 2 3 1 3 1 ...
##  $ race         : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capitalgain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capitalloss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hoursperweek : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ nativecountry: Factor w/ 8 levels "misLevel","SEAsia",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ income       : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Summarize all data sets

```
##       age                workclass        fnlwgt
##  Min.   :17.00   Private      :22696   Min.   :  12285
##  1st Qu.:28.00   SelfEmpInc   : 2541   1st Qu.: 117827
##  Median :37.00   LocGov       : 2093   Median : 178356
##  Mean   :38.58   misLevel     : 1836   Mean   : 189778
##  3rd Qu.:48.00   StateGov     : 1298   3rd Qu.: 237051
##  Max.   :90.00   SelfEmpNotInc: 1116   Max.   :1484705
##                  (Other)      :  981
##       education      educationnum     maritalstatus
##  graduate  :12646   Min.   : 1.00   divorce   : 5468
##  highsch   : 3896   1st Qu.: 9.00   married   :15394
##  master    : 1723   Median :10.00   notmarried:10683
##  secndrysch: 1608   Mean   :10.08   widowed   :  993
##  upperprim :  646   3rd Qu.:12.00   NA's      :   23
##  (Other)   :  965   Max.   :16.00
##  NA's      :11077
##         occupation       relationship                   race
##  misLevel    : 1843   husband   :13193   Amer-Indian-Eskimo:  311
##  clerical    : 3770   wife      : 1568   Asian-Pac-Islander: 1039
##  lowskillabr :15555   outofamily: 8305   Black             : 3124
##  highskillabr: 6184   unmarried : 3446   Other             :  271
##  agricultr   :  994   relative  :  981   White             :27816
##  NA's        : 4215   ownchild  : 5068
##
##      sex         capitalgain     capitalloss      hoursperweek
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##       nativecountry     income
##  NorthAmerica:30555   <=50K:24720
##  misLevel    :  663   >50K : 7841
##  Europe      :  492
##  Asia        :  467
##  SouthAmerica:  137
##  (Other)     :  227
##  NA's        :   20
```

cleaning data with NAs

We again see that independent variables education,maritalstatus,occupation,nativecountry have 11077,23 ,4215,20 missing value respecively. Here I imputed missed values using missForest.

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
```

split the data into 75:25 ratio.

(i). Validation Set Approach (VSA)

significant predictors are age, workclassSelfEmpInc,fnlwgt,educationnum and maritalstatusmarried. As for the statistical significant variables, age and educationnum has the lowest p value suggesting a strong association with the response, income

```
## [1] 0.1793391
```

KNN

```
## $K
## [1] 100
##
## $misclass
##   [1] 0.3131065 0.3244073 0.2897678 0.2951726 0.2805552 0.2795725 0.2696229
##   [8] 0.2710969 0.2620071 0.2611473 0.2594276 0.2577079 0.2550055 0.2525488
##  [15] 0.2509520 0.2498465 0.2505835 0.2513205 0.2473898 0.2478811 0.2468984
##  [22] 0.2476354 0.2462842 0.2462842 0.2454244 0.2467756 0.2456701 0.2448102
##  [29] 0.2445645 0.2443189 0.2432134 0.2433362 0.2428449 0.2435819 0.2437047
##  [36] 0.2439504 0.2440732 0.2441960 0.2437047 0.2435819 0.2435819 0.2434590
##  [43] 0.2433362 0.2435819 0.2435819 0.2433362 0.2433362 0.2433362 0.2433362
##  [50] 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362
##  [57] 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362
##  [64] 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362
##  [71] 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362
##  [78] 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362
##  [85] 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362
##  [92] 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362 0.2433362
##  [99] 0.2433362 0.2433362
##
## $Kmin
## [1] 33
```

At k=33 , knn gives less miss classification error

```
## [1] 0.242722
```

LDA

```
## [1] 0.1770053
```

QDA

QDA did not accept catogorical explanatory variables. I think qda assumes real values (and not factors) in the explanatory variables. Hence I removemed these variables from fromula to run the QDA function. (+ workclass + maritalstatus)

```
## [1] 0.2154526
```

Mclust

```
## [1] 0.2410023
```

```
## [1] 0.2067314
```

LOOCV Approach

GLM

I Used the for loop to split the data in 1:n-1 ratio.

```
## [1] 0.1873464
```

LDA

I removed + maritalstatus+ workclass variables from the model because it was shoiwing , Error in lda.default(x, grouping, …) : variable 14 appears to be constant within groups I tried to explore about this error but could reach final conculusion about why this error show up.

```
## [1] 0.2168305
```

QDA-LOOCV

```
## [1] 0.213145
```

KNN-LOOCV

```
## [1] 0.3667076
```

mclust-loocv

5-Fold

glm

```
## [1] 0.1219478
```

KNN (choose the best of K)

```
## [1] 0.2909104
```

LDA

```
## [1] 0.2194654
```

QDA

```
## [1] 0.2096682
```

MclustDA - best model chosen by BIC

```
## $error
## [1] 0.2651515
```

MclustDA with modelType="EDDA

```
## $error
## [1] 0.202498
```

Fit a Support Vector Machine (SVM) classification model

Validation Set Approach (VSA)

```
## [1] 0.0002456701
```

LOOCV

```
##
## Error estimation of 'svm' using leave-one-out: 0.001228501
```

5-fold Cross Validation

```
##
## Error estimation of 'svm' using 5-fold cross validation: 0.0001638002
```