IDS 575 Homework 1

Submitted by

Chaitanya Muppala 662043907
CJ All 650065604
Manisha Singh 674146073

Problem 1: Instance-based Learning

Restaurant	Food?	Ambiance?	Service?	Like?
1	good	normal	fair	yes
2	average	superior	bad	yes
3	poor	superior	bad	no
4	good	inferior	fair	no
5	average	normal	bad	no

a) Define each attribute F(ood), A(mbiance), S(ervice), and the output space L(ike), respectively as a set of possible values. What is the size |X| of the instance space X?

Answer: The instance space is also known as input space. Here Food attribute has 3 possible values, Ambiance has 3 values and service has 2. The total possible combinations of these three attributes is the size of instance space. The size of the instance space X is 3*3*2 = 18

b) Hypothesis space is defined as a set of all possible function h : F × A × S → L. What is the size |H| of the hypothesis space H?

Answer: Hypothesis is the total number of possibilities h : F × A × S → L. From above, we know that the size of instance space is 18 , and has a possibility of two output values (Like – Yes or No). Then the size of the hypothesis space H is 2^18 = 262144

c) Initially Juno thought that she will like any restaurant that serves good food. Is this hypothesis h consistent with the training set D? Why or why not?

Answer: For function -> {F,A,S}This hypothesis is not consistent with the training data D because of restaurant 4. If you observe restaurant 4, the food is rated good and still Juno didn't like this restaurant.

d) (+5 pts) Count the number of hypotheses in H that are consistent to her training data D.

Answer: For hypothesis space f ->{F,A} 5 of its instances matches with the training data. Five of the hypotheses for f ->{F,A} consistent with the training data.

For {A,S} -> 4 are consistent, {F,S} -> 3 are consistent . Total = 5+4+3 = 12

(3)(a)	Food	Ambiance	Service	Like
1	good (1)	normal (0)	fair (1)	yes
2	average (0)	superior (1)	bad (-1)	yes
3	poor (-1)	superior (1)	bad (-1)	no
4	good (1)	inferior (-1)	fair (1)	no
5	average (0)	normal (0)	bad (-1)	no

{F,A}	{F,S}	{A,S}	Like
1 Yes	2 Yes	1 Yes	
1 Yes	-1 No	0 No	"
0 No	-2 No	0 No	"
0 No	2 Yes	0 No	
0 No	-1 No	-1 No	

{F,A} matches with the training data ... consistent all five values.

e) Let H0 be the restricted hypothesis space (H0 ⊆ H) that satisfies the above numeric prediction rule. Measure the size |H0 | of the new hypothesis space H0 . (Hint: Consider each of the three cases where only {F, A}, {F, S}, or {A, S} matters and the other one attribute doesn't care. Make sure that H0 is a subset of H)

Answer: We have about 262144 total hypotheses, which is too large. To restrict the hypothesis, we are using the sum of two attribute values and we are considering these three cases {F, A}, {F, S}, {A,S}.

For {F,A} , we have Instances of 3*3 =9 -> hypothesis 2^9 = 512

{F,S} -> 3*2 = 6 -> 2^6 =64

{A,S} -> 3*2 =6 -> 2^6 = 64

The size of restricted hypothesis space {F,A}+{F,S}+{A,S} = 512+64+64 = 640. The restricted hypothesis is a subset of total hypothesis.

Problem 2: k-Nearest Neighbors Algorithm

a) **Draw the decision boundaries of D1 and D2 when k = 1.**

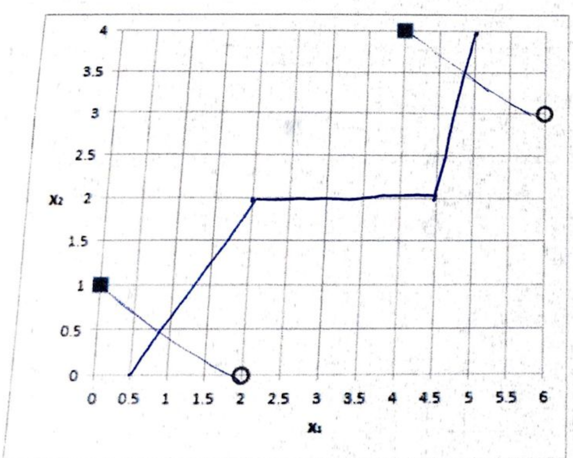


Figure 1: The training set S_1

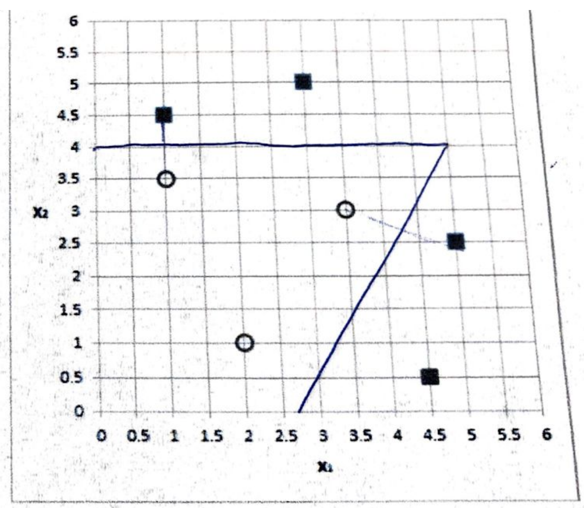


Figure 2: The training set S_2

b) **Which label would you suggest for (3, 2) and (4, 2) in D1? Ties are broken by predicting the positive class.**

For (3,2) we suggest a positive/white label. For (4,2), we suggest a negative/black label.

2b. $B_1 = (4,4)$ $W_1 = (6,3)$
$B_2 = (0,1)$ $W_2 = (2,0)$

For $P = (3,2)$:
$d(P, B_1) = \sqrt{5}$
$d(P, B_2) = \sqrt{10}$ $\rightarrow$ tie $\rightarrow$ white
$d(P, W_1) = \sqrt{10}$
$d(P, W_2) = \sqrt{5}$

For $P = (4,2)$:
$d(P, B_1) = \sqrt{4}$

c) Which label would you suggest for (4.5, 4) and (4, 2) be in D2? Ties are broken by predicting the positive class.

For (4.5,4) and (4,2), we suggest positive/white labels.

2C. $B_1 = (3,5)$ $W_1 = (3.5, 3)$
$B_2 = (5, 2.5)$
$B_3 = (4.5, 0.5)$

For $P = (4.5, 4)$:
$d(P, B_1) = \sqrt{3.25}$
$d(P, B_2) = \sqrt{2.5}$ → white
$d(P, B_3) = \sqrt{12.25}$
$d(P, W_1) = \sqrt{2}$

$B_1 = (4.5, 0.5)$ $W_1 = (2, 1)$
$B_2 = (5, 2.5)$ $W_2 = (3.5, 3)$

For $P = (4, 2)$:
$d(P, B_1) = \sqrt{2.5}$
$d(P, B_2) = \sqrt{1.25}$ → white
$d(P, W_1) = \sqrt{5}$
$d(P, W_2) = \sqrt{1.25}$

d) When k > 1, a partition of spaces like the above is called the k th-order Voronoi Diagram or Voronoi Tesselation. Try to draw the decision boundaries of D1 when k = 3. (Hint: Try to draw every bisector between all pairs of positive and negative examples)

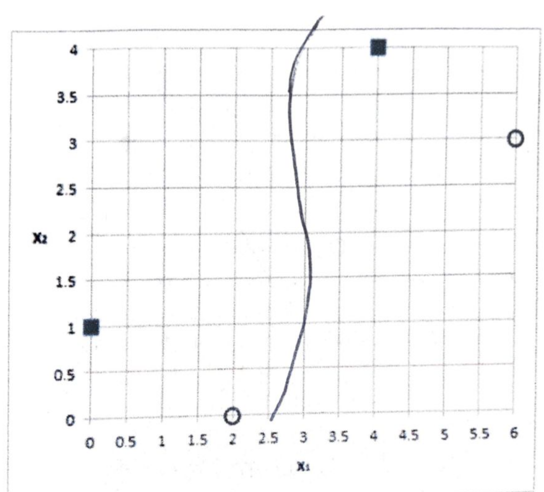
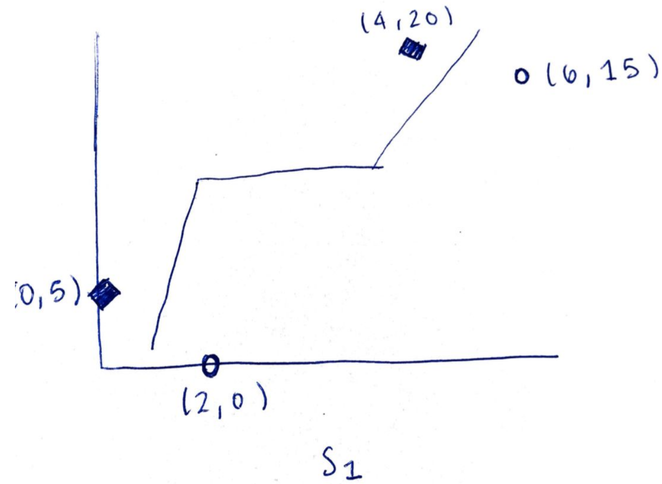


Figure 1: The training set S_1

e) **(+ 5pts) If the x2-coordinate of all four example points in D1 were multiplied by 5, what would happen to its decision boundary when k = 1? Draw another picture. Could this change cause any problem when working with real data?**



Problem 3: Multivariate Calculus

Evaluate individual sub-problems including all of your derivations. Justify your solutions if a problem requires a mathematical proof.

a. **Say g(z) = 1 1+e−z . Prove g 0 (z) = g(z)(1 − g(z)).**

① $\frac{d}{dz}\left(\frac{1}{e^{-z}+1}\right)$:

chain rule - Let $u = e^{-z}+1$

$$\frac{d}{du}\left(\frac{1}{u}\right) = -\frac{1}{u^2}$$

$$= \frac{\frac{d}{dz}(1+e^{-z})}{(1+e^{-z})^2}$$

$$= \frac{1}{(1+e^{-z})^2}\left(\frac{d}{dz}(1)+\frac{d}{dz}(e^{-z})\right)$$

$$= \frac{1}{(1+e^{-z})^2}\left(0+\frac{d}{dz}(e^{-z})\right)$$

$$= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\left(\frac{d}{dz}(+z)\right)\right)$$

$$= \frac{1}{(1+e^{-z})^2}(e^{-z})$$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

② $g'(z) = g(z)(1-g(z))$

$$= \frac{1}{1+e^{-z}}\left(1-\frac{1}{1+e^{-z}}\right)$$

$$= \frac{1}{1+e^{-z}} - \frac{1}{(1+e^{-z})^2}$$

$$= \frac{1+e^{-z}-1}{(1+e^{-z})^2}$$

$$= \frac{e^{-z}}{(1+e^{-z})^2} \checkmark$$

b. **Say $h(x_1, x_2) = (\theta_0 + \theta_1 x_1 + \theta_2 x_2^2)^2$. Evaluate $\frac{\partial h}{\partial x_1}$ and $\frac{\partial h}{\partial x_2}$.**

(2) $(a+b+c)^2 =$

$a^2+b^2+c^2+2ab+2bc+2ca$ } Using product rule $ax^b = abx^{b-1}$

$(\theta_0+\theta_1x_1+\theta_2x_2)^2$

$= \theta_0^2 + \theta_1^2 x_1^2 + \theta_2^2 x_2^2 + 2\theta_0\theta_1x_1 + 2\theta_1x_1\theta_2x_2$
$+ 2\theta_0\theta_2x_2$

$\frac{\partial h}{\partial x_1} = 0 + \theta_1^2 2x_1 + 0 + 2\theta_0\theta_1 + 2\theta_1\theta_2x_2$
$+ 0$

$\frac{\partial h}{\partial x_2} = 0 + 0 + \theta_2^2 2x_2 + 0 + 2\theta_1\theta_2x_1$
$+ 2\theta_0\theta_2$

$$\boxed{\begin{aligned}\frac{\partial h}{\partial x_1} &= \theta_1^2 2x_1 + 2\theta_0\theta_1 + 2\theta_1\theta_2x_2 \\ \frac{\partial h}{\partial x_2} &= \theta_2^2 2x_2 + 2\theta_1\theta_2x_1 + 2\theta_0\theta_2\end{aligned}}$$

c. **Say $J(x, y) = y \log(\theta_0 + \theta_1 x) + (1 - y)\log(1 - \theta_0 - \theta_1 x)$. Evaluate $\frac{\partial J}{\partial \theta_0}$ and $\frac{\partial J}{\partial \theta_1}$.**

$$\frac{\partial J}{\partial \theta_0} = \frac{y - \theta_0 - \theta_1 x}{(\theta_0+\theta_1 x)(1-\theta_0-\theta_1 x)}$$

$$\frac{\partial J}{\partial \theta_1} = \frac{x(\theta_0+\theta_1 x - y)}{(\theta_0+\theta_1 x)(1-\theta_0-\theta_1 x)}$$

$$\frac{\partial J}{\partial \theta_0} = (1-y)\frac{\partial}{\partial \theta_0}(\log(1-\theta_0-\theta_1 x) + y\left(\frac{\partial}{\partial \theta_0}(\log(\theta_0+\theta_1 x))\right)$$

$\left\{\begin{array}{l}\text{use chain rule,}\\ u = -\theta_0 - \theta_1 x + 1\\ \frac{\partial}{\partial u}(\log(u)) = 1/u\end{array}\right.$

$$= y\left(\frac{\partial}{\partial \theta_0}(\log(\theta_0+\theta_1 x))\right) + (1-y)\frac{\frac{\partial}{\partial \theta_0}(1-\theta_0-\theta_1 x)}{1-\theta_0-\theta_1 x}$$

$$= y\left(\frac{\partial}{\partial \theta_0}(\log(\theta_0+\theta_1 x))\right) + (1-y)\frac{-\left(\frac{\partial}{\partial \theta_0}(\theta_0) + \frac{\partial}{\partial \theta_0}(-\theta_1 x)\right)}{1-\theta_0-\theta_1 x}$$

$$= \frac{(1-y)\left(-\left(\frac{\partial}{\partial \theta_0}\theta_0 + \frac{\partial}{\partial \theta_0}(-\theta_1 x)\right)\right)}{1-\theta_0-\theta_1 x} + y\frac{\partial}{\partial \theta_0}\log(\theta_0+\theta_1 x)$$

$$= -\frac{1-y}{1-\theta_0-\theta_1 x} + y\left(\frac{\partial}{\partial \theta_0}\log(\theta_0+\theta_1 x)\right)$$

$\left\{\begin{array}{l}\text{use chain rule,}\\ u = \theta_0 + \theta_1 x\\ \frac{\partial}{\partial u}\log u = 1/u\end{array}\right.$

$$= -\frac{1-y}{1-\theta_0-\theta_1 x} + \frac{\frac{\partial}{\partial \theta_0}(\theta_0+\theta_1 x)}{\theta_0+\theta_1 x}y$$

$$= -\frac{1-y}{1-\theta_0-\theta_1 x} + \frac{y}{\theta_0+\theta_1 x}$$

$$= \boxed{\frac{y-\theta_0-\theta_1 x}{(\theta_0+\theta_1 x)(1-\theta_0-\theta_1 x)}}$$

$$\frac{\partial J}{\partial \theta_1} = (1-y)\left(\frac{\partial}{\partial \theta_1}\log(1-\theta_0-\theta_1 x) + y\left(\frac{\partial}{\partial \theta_1}\log(\theta_0+\theta_1 x)\right)\right)$$

$\left\{\begin{array}{l}\text{use chain rule,}\\ u = 1-\theta_0-\theta_1 x\end{array}\right.$

$$= y\left(\frac{\partial}{\partial \theta_1}\log(\theta_0+\theta_1 x)\right) + (1-y)\frac{\frac{\partial}{\partial \theta_1}(1-\theta_0-\theta_1 x)}{1-\theta_0-\theta_1 x}$$

$$= \frac{(1-y)\left(\frac{\partial}{\partial \theta_1}(-\theta_0) - x\left(\frac{\partial}{\partial \theta_1}\theta_1\right)\right)}{1-\theta_0-\theta_1 x} + y\left(\frac{\partial}{\partial \theta_1}\log(\theta_0+\theta_1 x)\right)$$

$$= y\left(\frac{\partial}{\partial \theta_1}\log(\theta_0+\theta_1 x)\right) - \frac{x(1-y)}{1-\theta_0-\theta_1 x}$$

$\left\{\begin{array}{l}\text{use chain rule,}\\ u = \theta_0 + \theta_1 x\end{array}\right.$

$$= \frac{x(1-y)}{1-\theta_0-\theta_1 x} + y\frac{\frac{\partial}{\partial \theta_1}(\theta_0+\theta_1 x)}{\theta_0+\theta_1 x}$$

$$= \frac{-x(1-y)}{1-\theta_0-\theta_1 x} + \frac{xy}{\theta_0+\theta_1 x}$$

$$= \boxed{\frac{x(\theta_0+\theta_1 x - y)}{(1-\theta_0-\theta_1 x)(\theta_0+\theta_1 x)}}$$

d. Given |X| = 10, |Y | = 3, count the number of mathematical functions f that maps each element in X to a element in Y .

For the function f, every element in X is mapped to exactly one element in Y. There are exactly three ways for an element of X to map to an element in Y, so there are 3^{10} (or 59,049) functions that map each X to Y.

e. For real vectors θ ∈ R n and x ∈ R n , hθ(x) = θ1x1 + ... + θnxn. Evaluate the gradient ∇xhθ(x) (with respect to x).

(3)(c)

$\theta \in R^n, \quad x \in R^n$

$h_\theta(x) = \theta_1 x_1 + \cdots + \theta_n x_n$

$\nabla_x h_\theta(x)$

$$= \left(\frac{d}{dx_1} h_\theta(x), \frac{d}{dx_2} h_\theta(x), \frac{d}{dx_3} h_\theta(x), \ldots, \frac{d}{dx_n} h_\theta(x)\right)$$

$$= (\theta_1, \theta_2, \theta_3 \cdots \theta_n)$$

$\nabla_x h_\theta(x)$ w.r.t x_i (from piazza hint).

$$= \frac{d}{dx_i} h_\theta(x)$$

$$= \frac{d}{dx_i} (\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_i \cdot x_i + \cdots \theta_n x_n)$$

sumrule

$$= \frac{d}{dx_i}(\theta_1 x_1) + \frac{d}{dx_i}(\theta_2 x_2) + \cdots + \frac{d}{dx_i}(\theta_i x_i) + \cdots$$

$$\cdots + \frac{d}{dx_i}(\theta_n x_n)$$

$$= 0 + 0 + 0 + \cdots + \theta_i + \cdots + 0$$

$$= \theta_i$$

f. **L2-norm measures a magnitude of a vector. In other words, kxk2 : R n −→ R is a function that maps a vector argument x ∈ R n into a scalar value. Evaluate the gradient ∇xkxk 2 2 if it is feasible. Argue why if it is not.**

Norm definition : https://en.wikipedia.org/wiki/Norm_(mathematics)#Euclidean_norm

③ f Lp-norm is defined as follows

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/2}$$

here we are using L2-norm

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$$

$$\|x\|_2^2 = \left(\left(\sum_{i=1}^{n} x_i^2\right)^{1/2}\right)^2$$

$$f(x) = \|x\|_2^2 = \sum_{i=1}^{n} x_i^2$$

gradient:

$$\frac{\partial (f(x))}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\sum_{i=1}^{n} x_i^2\right)$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial x_j}(x_i^2)$$

$i \neq j \quad = 0$

otherwise $\quad = 2x_i$

$$\nabla_x \|x\|_2^2 = 2x$$

Problem 4: Linear Regression

a. **Take a look at our data:**

   i. **What is the number of training examples m and the number of features n except the name attribute?**

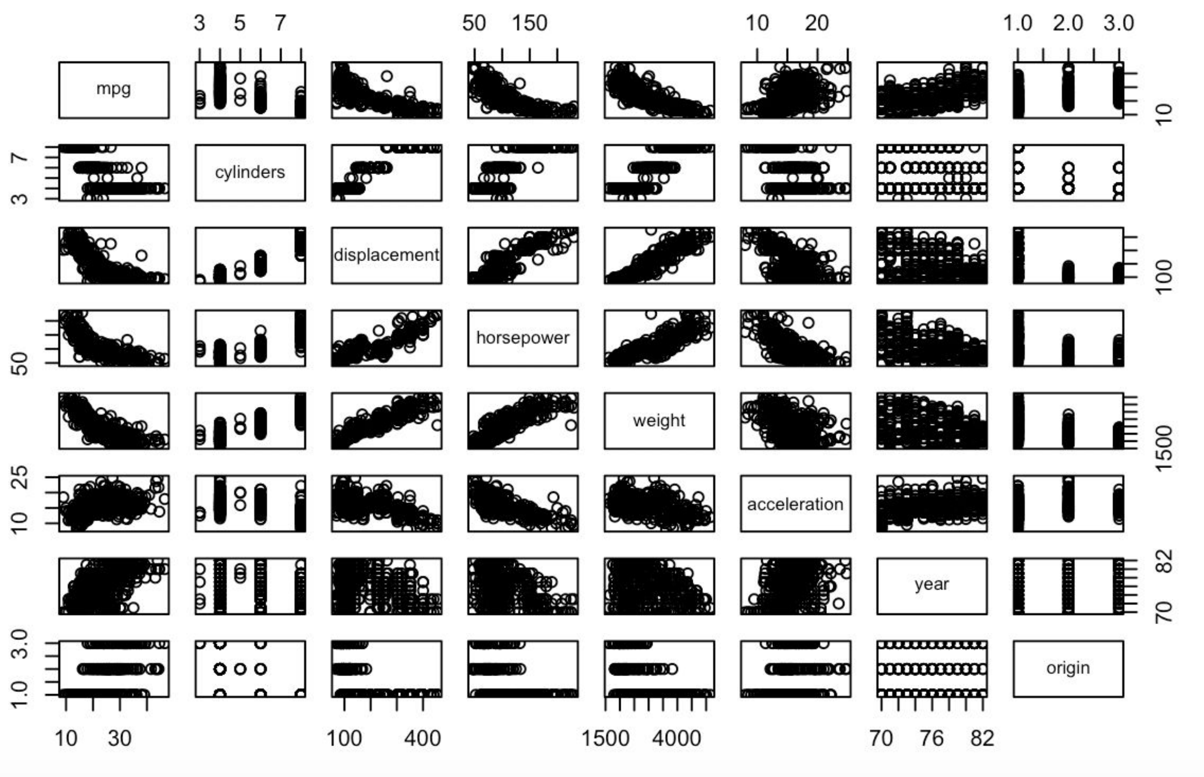
      When looking at the dataset, we see that there are m=392 training examples/observations, and that there are n=8 features excluding "name" which include "mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "year", and "origin".

   ii. **Thinking this data as a matrix X ∈ R m×n , is X a skinny/tall matrix or fat/wide matrix?**

      Since this matrix is 392 x 8 (meaning there are more rows than columns) it is a skinny/tall matrix.

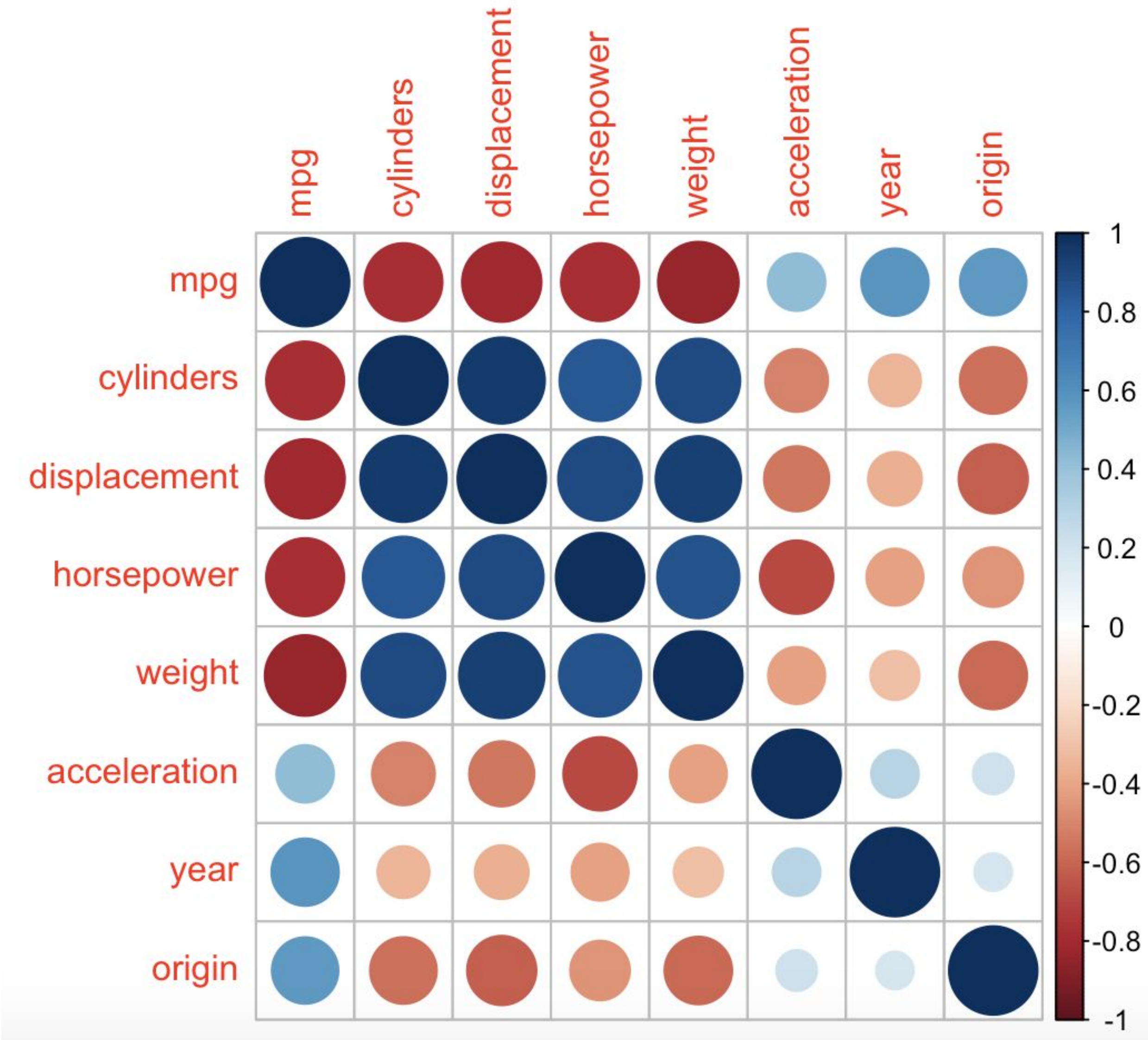
b. **Perform the basic exploratory analysis by computing and visualizing correlations:**

   i. **Draw the plot of correlations between every pair of features:**



Auto = select(Auto, mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin)
plot(Auto)

   ii. **Which features are highly correlated to one another?**



correlations = cor(Auto)
corrplot(correlations)

From the two visuals shown above, we can safely infer that variables such as displacement and cylinders are highly correlated. Weight is highly positively correlated with cylinders and displacement as well.

c. **Perform linear regression by putting mpg as an output variable based on all other features except the name attribute:**

```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin, data = Auto)
Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate       Std. Error   t value  Pr(>|t|)
(Intercept)  -17.218435     4.644294  -3.707  0.00024 ***
cylinders      -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight         -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year            0.750773   0.050973  14.729  < 2e-16 ***
origin          1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,  Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i. **Is there any relationship between the input features and the output response?**

   Yes. The linear regression model for the Auto dataset is shown to be predictive because of each feature, with an adjusted r squared of 81.82% and a p value of under 2.2E-16 for the model. 81.82% of the variance in mpg is explained by the predictors in the model.

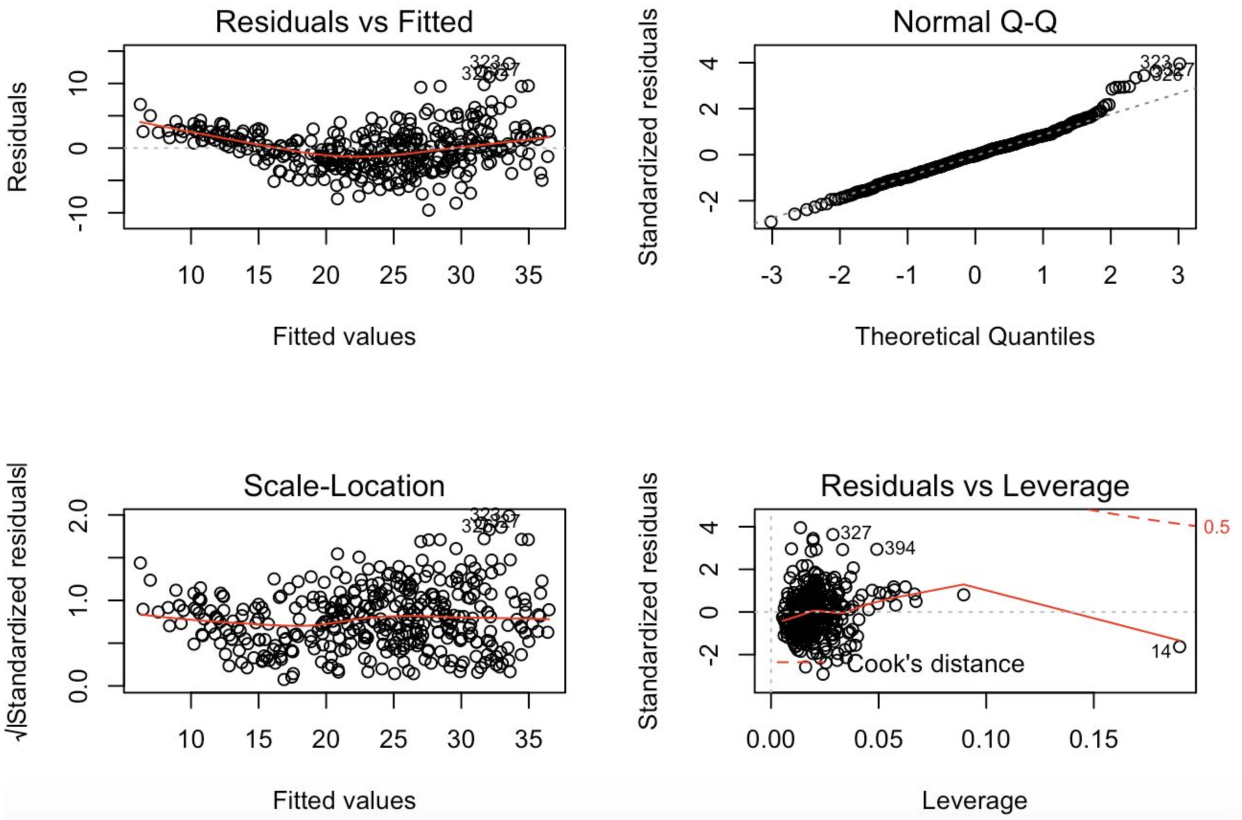
ii. **Which features appear to have a statistically significant relationship to the output response?**

   Weight and Year. These are the only two significant variables at the 5% level because both features are predictive at a less than .01% level. They both also have p-values at <2e-16 which nearly match the model p-value of 2.2E-16.

iii. **What does the coefficient for the year variable suggest?**

The value of 0.750773 means that there’s a positive slope and it’s a positive correlation. What does this mean? Newer cars should have a higher mpg.

d. Produce diagnostic plots of the linear regression fit by using plot:



par(mfrow=c(2,2))
plot(lmMpg)

i. **Any problems in the fit?**

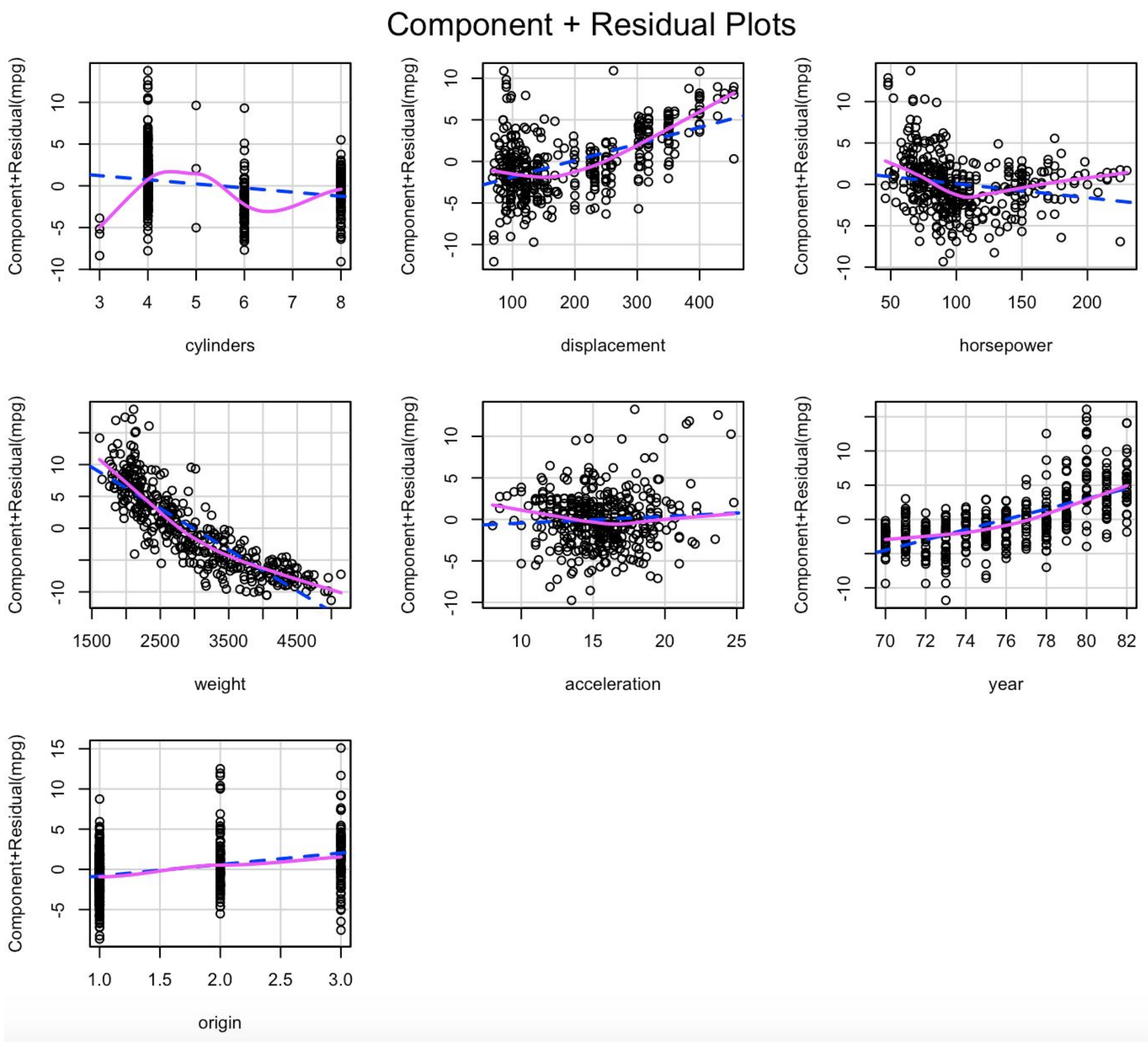
Yes. The residual values for the fit aren’t normally distributed, there are multiple outliers within the graphs, and there is nonlinearity within the model.

Normally, a Q-Q plot would be a 45 degree line if the residual values are normally distributed with a mean of zero. The plot indicates that this isn’t true, and that it’s angle is slightly lower instead going just under 45 degrees.

If the dependent variables are linearly related to the independent variables, then the Residuals vs Fitted Values chart should show no systematic relationship between the fitted values and the residuals. The shape of the graph indicates that this is not the case.

We can further augment the argument of nonlinearity being present by using component + residual plots. The graph for mpg and origin is misleading since origin should be categorical.

ii. **Do the residual plots suggest any unusually large outliers?**
Yes. There are numerous unusually large outliers according to the Residual vs Fitted, Scale-Location, and Residuals vs Leverage graphs. In particular, we see points 323, 325, 327, and 394 on the graphs. We can also see more on the component + residual plots as well:



coefplot(lmMpg)
crPlots(lmMpg)

e. **Try a few different transformations of the variables, such as log x_j , $\sqrt{x_j}$, x^2_j . Can you make a better fit?**

i. Taking the Log of each variable and putting it into an lm model

```
lm(formula = mpg ~ Logcyl + Logdis + Loghorse + Logweight + Logacc +
   Logyear, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5641 -1.7873 -0.0611  1.5810 13.2714 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -62.413     17.650  -3.536 0.000456 ***
Logcyl         2.750      1.626   1.691 0.091585 .  
Logdis        -3.406      1.355  -2.513 0.012371 *  
Loghorse      -6.386      1.563  -4.085 5.36e-05 ***
Logweight    -11.905      2.240  -5.316 1.80e-07 ***
Logacc        -5.326      1.622  -3.283 0.001119 ** 
Logyear       54.825      3.595  15.250  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.103 on 385 degrees of freedom
Multiple R-squared:  0.8444,  Adjusted R-squared:  0.8419 
F-statistic: 348.1 on 6 and 385 DF,  p-value: < 2.2e-16
```

It appears that the multiple and adjusted r squared values increase, while the p-value remains constant. This means that we do get a better fit if we were to take the logarithmic value for each variable.

ii. Taking the Sqrt of each variable and putting it into an lm model

```
Call:
lm(formula = mpg ~ Sqrcyl + Sqrdis + Sqrhorse + Sqrweight + Sqracc +
   Sqryear, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.0770 -1.9915 -0.2719  1.7993 13.9583 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -45.0956     9.3107  -4.843 1.85e-06 ***
```

Sqrcyl	1.0224	1.5417	0.663	0.5076	
Sqrdis	-0.1794	0.2132	-0.841	0.4007	
Sqrhorse	-0.5345	0.3090	-1.730	0.0845	.
Sqrweight	-0.6222	0.0807	-7.709	1.09e-13	***
Sqracc	-0.9155	0.8524	-1.074	0.2835	
Sqryear	12.7588	0.8777	14.537	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.281 on 385 degrees of freedom
Multiple R-squared: 0.826,   Adjusted R-squared: 0.8233
F-statistic: 304.7 on 6 and 385 DF, p-value: < 2.2e-16

The r squared value also increases for this column while the p-value still remains constant. It appears that the model has a better fit due to the higher r squared value.