Week 7

ASSIGNMENT 7

Load the dataset iris.csv as a dataframe **iris_data**, with the first column as index headers, first row as column headers, dependent variable as factor variable, and answer the following questions.

The iris dataset contains four Sepal and Petal features (Sepal Length, Sepal Width, Petal Length, Petal Width, all in cm) of 50 equal samples of 3 different species of the iris flower (Setosa, Versicolor, and Virginica).

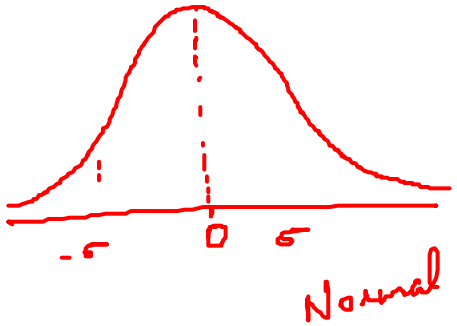
1) What are the dimensions of the dataframe?

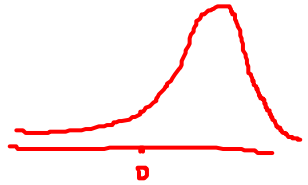
- ○ (150, 5)
- ○ (150, 4)
- ○ (50, 5)
- ○ None of the above

```
> iris_data = read.csv("/home/gitaa/Downloads/iris.csv", header = TRUE, row.names = "ID", stringsAsFactors=TRUE)
> dim(iris_data)
[1] 150   5
```

2) What can you comment on the distribution of the independent variables in the dataframe?

- All the variables are normally distributed
- The variables Sepal Length and Sepal Width are not normally distributed
- The variable Petal Length alone is normally distributed
- ✓ None of the above





→ plot

→ mean & median

```
> summary(iris_data)
  SepalLength      SepalWidth     PetalLength      PetalWidth
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

3) How many rows in the dataset contain missing values?

- 10
- 5
- 25
- 0

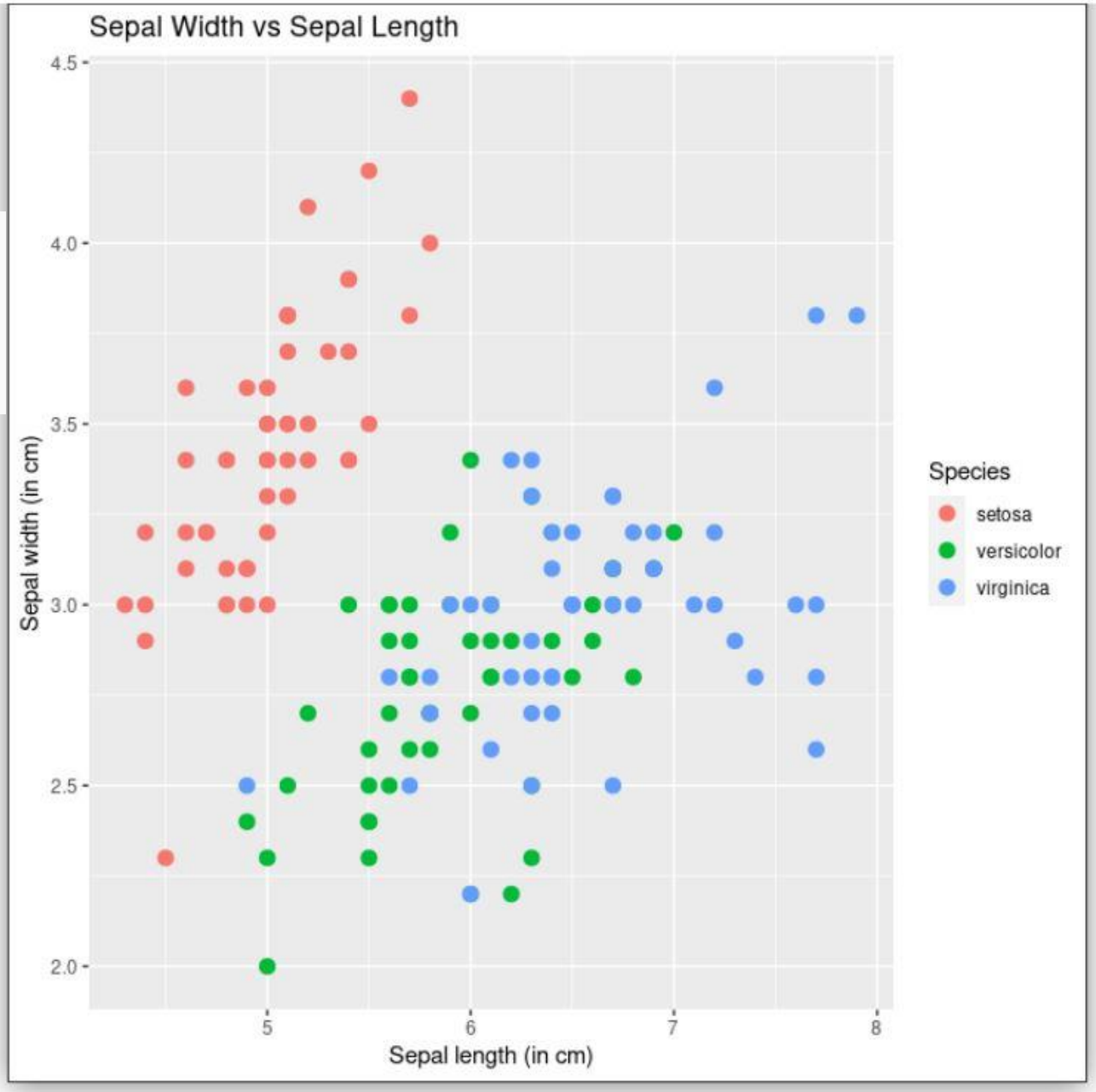
is na

is not applicable

```
> sum(is.na(iris_data))
[1] 0
```

4) What can be interpreted from the plot shown below?

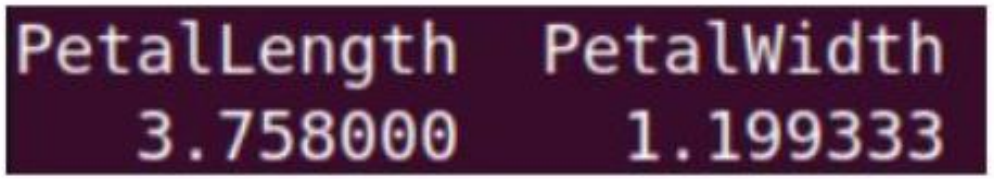
- [ ] Sepal widths of Versicolor flowers are lesser than 3 cm
- [ ] Sepal lengths of Setosa flowers are lesser than 6 cm
- [ ] Sepal lengths of Virginica flowers are greater than 6 cm
- [ ] Sepals of Setosa flowers are relatively more wider than Versicolor flowers



Solution: b) d)

Feedback: Based on the given plot, it can be seen that Setosa flowers have wider sepals compared to Versicolor and Virginica flowers. Also, Sepal lengths of Setosa flowers are less than 6 cm.

5) Which of the following code blocks can be used to summarize the data (finding the mean of the columns PetalLength and PetalWidth), similar to the one given below.



- [x] lapply(iris_data[, 3:4], mean)
- [ ] sapply(iris_data[, 3:4], 2, mean)
- [x] apply(iris_data[, 3:4], 2, mean)
- [ ] apply(iris_data[, 3:4], 1, mean)

Solution: a) c)

Feedback: Option d) gives the mean value across the columns for each and every row in the dataframe.

```
> apply(iris_data[, 3:4], 1, mean)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
0.80 0.80 0.75 0.85 0.80 1.05 0.85 0.85 0.80 0.80 0.85 0.90 0.75 0.60 0.70 0.95
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
0.85 0.85 1.00 0.90 0.95 0.95 0.60 1.10 1.05 0.90 1.00 0.85 0.80 0.90 0.90 0.95
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
0.80 0.80 0.85 0.70 0.75 0.75 0.75 0.85 0.80 0.80 0.75 1.10 1.15 0.85 0.90 0.80
  49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64
0.85 0.80 3.05 3.00 3.20 2.65 3.05 2.90 3.15 2.15 2.95 2.65 2.25 2.85 2.50 3.05
  65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80
2.45 2.90 3.00 2.55 3.00 2.50 3.30 2.65 3.20 2.95 2.80 2.90 3.10 3.35 3.00 2.25
  81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96
2.45 2.35 2.55 3.35 3.00 3.05 3.10 2.85 2.70 2.65 2.80 3.00 2.60 2.15 2.75 2.70
  97   98   99  100  101  102  103  104  105  106  107  108  109  110  111  112
2.75 2.80 2.05 2.70 4.25 3.50 4.00 3.70 4.00 4.35 3.10 4.05 3.80 4.30 3.55 3.60
 113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128
3.80 3.50 3.75 3.80 3.65 4.45 4.60 3.25 4.00 3.45 4.35 3.35 3.90 3.90 3.30 3.35
 129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144
3.85 3.70 4.00 4.20 3.90 3.30 3.50 4.20 4.00 3.65 3.30 3.75 4.00 3.70 3.50 4.10
 145  146  147  148  149  150
4.10 3.75 3.45 3.60 3.85 3.45
```

Option b) throws an error as the arguments are invalid.

```
> sapply(iris_data[, 3:4], 2, mean)
Error in match.fun(FUN) : '2' is not a function, character or symbol
```

Option a) and c) provide the output as follows

```
> lapply(iris_data[,3:4], mean)
$PetalLength
[1] 3.758

$PetalWidth
[1] 1.199333
```

```
> apply(iris_data[, 3:4], 2, mean)
PetalLength  PetalWidth
   3.758000    1.199333
```

6) Which of the following packages must be imported to use the logistic regression function ***glm()***?

- ROCR
- dplyr
- caTools
- None of the above

Solution: d)

Feedback: The logistic regression function ***glm()*** is an in-built function available with R Programming.

Encode the dependent variable **Species**. Split the given dataset **iris_data** into training and test sets in the ratio 7:3, with the seed value as 123. Apply the logistic regression model in the training set.

7) Which of the following parameters are significant with 95% confidence interval?

- [ ] Sepal Length
- [ ] Intercept
- [ ] Petal Width
- [ ] Petal Length

Species
Setosa
Versicolor
Virginica

7:3
seed = 123, 0, 1,
98%
98%

Solution: b) c) d)

Feedback:

```
> iris_data = read.csv("/home/gitaa/Downloads/iris.csv", header = TRUE, row.names = "ID", stringsAsFactors=TRUE)
> library(caTools)
> set.seed(123)
> split_data = sample.split(iris_data$Species, SplitRatio = 0.7)
> train_iris_data = subset(iris_data, split_data == TRUE)
> dim(train_iris_data)
[1] 105   5
> test_iris_data = subset(iris_data, split_data == FALSE)
> dim(test_iris_data)
[1] 45  5
```

```
> train_iris_data$Species <- as.numeric(train_iris_data$Species)
> test_iris_data$Species <- as.numeric(test_iris_data$Species)
```

```
> logit_model <- glm(Species ~ SepalLength + SepalWidth + PetalLength + PetalWidth, data = train_iris_data)
> summary(logit_model)

Call:
glm(formula = Species ~ SepalLength + SepalWidth + PetalLength +
    PetalWidth, data = train_iris_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.51042  -0.15239   0.01128   0.09432   0.50864

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.24047    0.24733   5.015 2.30e-06 ***
SepalLength  -0.04695    0.06596  -0.712   0.4782
SepalWidth   -0.13111    0.07029  -1.865   0.0651 .
PetalLength   0.16822    0.06423   2.619   0.0102 *
PetalWidth    0.66868    0.10637   6.286 8.58e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.04197323)

    Null deviance: 70.0000  on 104  degrees of freedom
Residual deviance:  4.1973  on 100  degrees of freedom
AIC: -28.072

Number of Fisher Scoring iterations: 2
```

From the given summary of the model applied on the training data, it can be observed that the parameters PetalWidth, PetalLength, and Intercept are significant variable as their p-value are lesser than the confidence level ($\alpha = 0.05$)

8) What is the coefficient of the variable Sepal Width in the fitted model?

- 1.24
- -0.15
- - 0.13
- 0.66

Solution: c)

Feedback: Based on the snippet for the previous question, the coefficient of the variable **SepalWidth** is -0.13111.

9) State whether the following statement is TRUE or FALSE.

Logistic Regression tends to overfit when we have a large number of independent variables present.

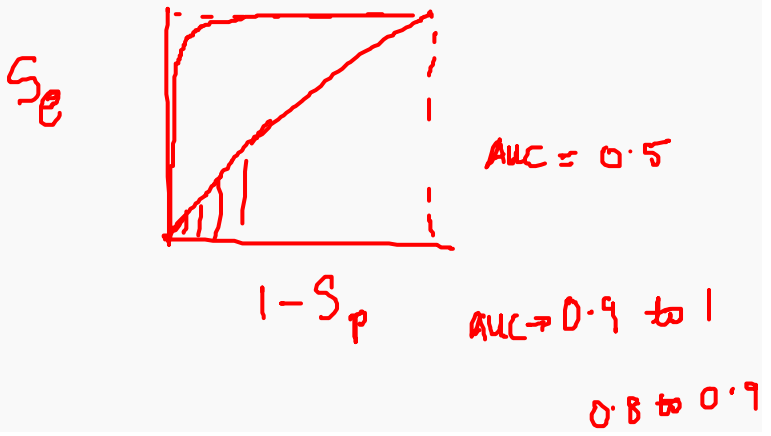
- True
- False

Solution: a)

Feedback: Yes, the logistic regression tends to overfit when we have a large number of independent variables present.

10) An ROC curve is plotted between.

- Sensitivity and Specificity
- Sensitivity and (1 – Specificity)
- (1 – Sensitivity) and Specificity
- None of the above



Solution: b)

Feedback: An ROC curve can be plotted between Sensitivity and (1 – Specificity) or True Positive Rate and False Positive Rate, to determine the better classifier that accurately predicts the classes.

Practice Questions

1) Which one of the following is called as the odds ratio?

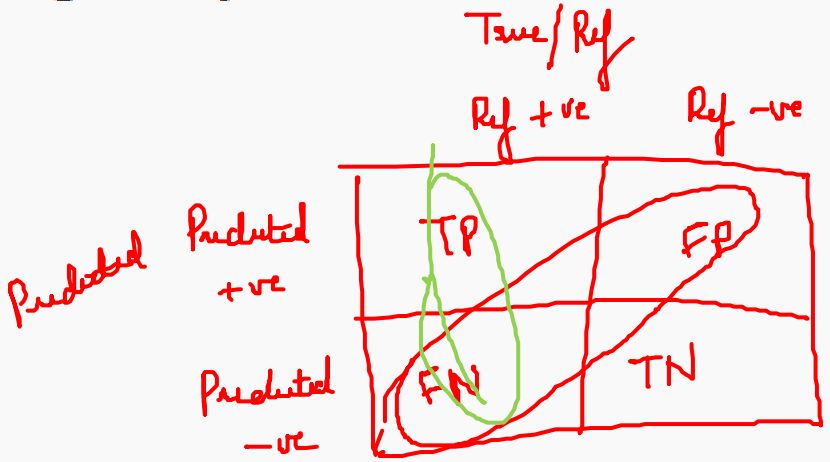
- [x] The ratio of the probability of an event occurring to the probability of the event not occurring
- [ ] The ratio of the probability of an event not occurring to the probability of the event occurring
- [ ] The probability of an event occurring
- [ ] The probability of an event not occurring

$p(x)$ $1-p(x)$

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

2) In confusion matrix, the misclassification rate is given by

- ✓ $\frac{False\ Negative + False\ positive}{Total\ number\ of\ samples}$
- ○ $\frac{True\ Negative + False\ positive}{Total\ number\ of\ samples}$
- ○ $\frac{False\ Negative + True\ positive}{Total\ number\ of\ samples}$
- ○ $\frac{True\ Negative + True\ positive}{Total\ number\ of\ samples}$



$$Acc = \frac{TP + TN}{N}$$

$$\text{Mis class. rate} = \frac{FN + FP}{N}$$

3) The value of both sensitivity and specificity lies between

- ○ -1 and 1
- ○ -1 and 0
- ○ -2 and 2
- ⊘ 0 and 1

$$S_e = \frac{TP}{TP + FN}$$

$$S_p = \frac{TN}{TN + FP}$$

4) In Logistic Regression a linear relationship is assumed between the independent variables and the

dependent

- ○ Sigmoid of the dependent variable
- ○ Log of the dependent variable
- ○ Sine of the dependent variable
- ✓ None of the above

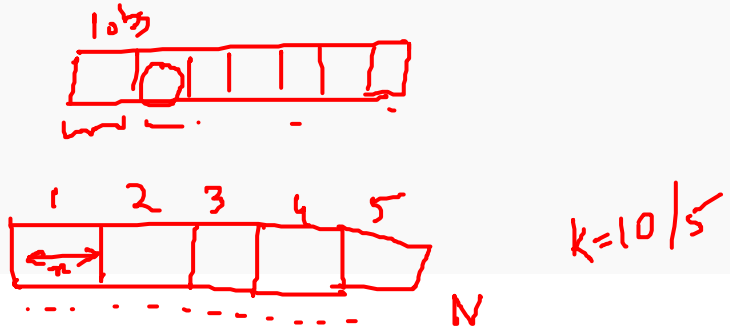
5) The confusion matrix for a binary classifier gives

- ○ Only True Positives and True Negatives
- ○ Only False Positives and False Negatives
- ○ Only True Positives, False Positives and False Negatives
- ✓ True Positives, False Positives, True Negatives and False Negatives

Previous course questions

1) Which among the following is not a type of cross-validation technique?

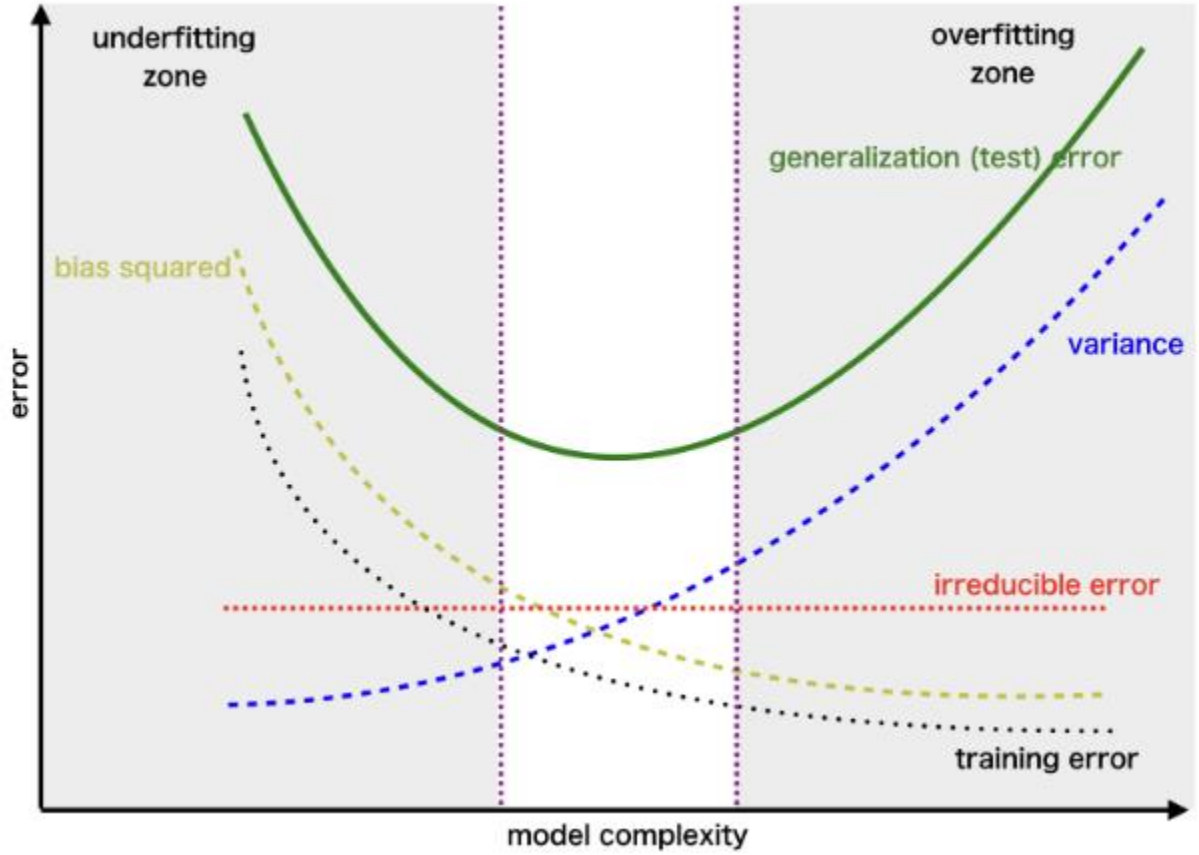
- LOOCV
- k-fold croos validation
- Validation set approach
- Bias variance trade off



- **LOOCV(Leave One Out Cross-Validation)** : is a type of cross-validation approach in which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set. In LOOCV, fitting of the model is done and predicting using one observation validation set.
- **k-fold Cross-validation** is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.
- **The Validation Set Approach** is a type of method that estimates a model error rate by holding out a subset of the data from the fitting process (creating a testing dataset). The model is then built using the other set of observations (the training dataset). Then the model result is applied on the testing dataset in which we can then calculate the error (testings dataset error). In summary, this general idea allows for the model to not overfit.
- **Bias Variance Trade off:** If the algorithm is too simple (hypothesis with linear equation) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree equation) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as a Trade-off or Bias Variance Trade-off. This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

Accepted Answers:
Bias variance trade off

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



2) Which among the following is a classification problem?

- [ ] Predicting the average rainfall in a given month.
- [x] Predicting whether a patient is diagnosed with a disease or not.
- [ ] Predicting the price of a house.
- [x] Predicting whether it will rain or not tomorrow.

Accepted Answers:
Predicting whether a patient is diagnosed with a disease or not.
Predicting whether it will rain or not tomorrow.

Consider the following confusion matrix for the classication of Hatchback and SUV:

		True	
		Hatchback	SUV
Prediction	Hatchback	55	5
	SUV	0	40

3) Find the accuracy of the model.

- 0.95
- 0.55
- 0.45
- 0.88

$N = 100$

$Acc = \frac{TP + TN}{N}$

$= \frac{55 + 40}{100}$

$= 0.95$

Accepted Answers:
0.95

4) Find the sensitivity of the model.

- 0.95
- 0.55
- 1
- 0.88

$$S_e = \frac{TP}{TP + FN} = 1$$

Accepted Answers:
1

5) Under the 'family' parameter of glm() function, which one of the following distributions correspond to logistic regression for a variable with binary output?

- Binomial
- Gaussian
- Gamma
- Poisson

Accepted Answers:
Binomial