CASE STUDY

Classification Problem

Submitted by: Raj Kumar, Msc Statistics (F)
(Practical-III, Gr A)

Due Date: 19th October, 2015

Date of Submission: 19th October, 2015

Supervisor's Remarks

Late Submission:

Plagiarism:

Completeness:

Quality of Content:

Results and Interpretations:

Additional Remarks:

About Classification Problem (CP)

The regression problem which we already know is called as a classification problem if the response is a discrete variable. In simpler words we want to classify an observation (univariate or multivariate) into one of several possible classes, or simply we want to estimate the probability given an observation that it belongs to one of the several possible classes. Due the fact that a discrete variable can't be normally distributed the application of linear regression becomes invalid as the assumption of normality of the observations is no longer satisfied. Indeed the response follows a multinomial distribution. Moreover, the expectation of response becomes uninterpretable in terms of a linear function of the features and also the variances do not remain constant across observations and hence causing heterosckedasticity. Hence such problems are outside the ambit of linear regression. There are several tools namely naïve Bayes classifier, logistic classifier, discriminant analysis, nearest neighbor approach, neural network etc. are available to be deployed in such situations. We will try some of them. The classification problems are quite naturally divided into two types:

1. Binary, where response has two possible classes meaning by an observation either belongs to a class or it doesn't, and

2. Multiclass, where the response has more than two possible classes.

Email Spam Filtering

Naïve Bayes Classifier: - The first task here is to classify whether an email is spam or non-spam (ham). An email will have some information in itself about the fact that its spam or not. We will find out that information and based on that we will estimate the probability of an email being spam. **An email is nothing but a text, which has words, symbols and numbers but all is text.** This entire text including the no. of persons to whom it's been sent, the name of the persons to whom it's been sent, the cc list, the bcc list (which we can't see though), the subject line, the body message (header, main message, the signature and the postscripts) will have information about the message being spam or non-spam. For this purpose we will make use of a technique called as the **Naïve Bayes Classifier**.

Email Spam Filtering using Naïve Bayes Classifier

Naïve Bayes classifier is a general technique for classification and can be used in the context of spam filtering. As we know the words of the message will tell us about it being spam or not, we can easily make a list of some common words which are generally seen in spam messages in a large frequency like congratulations, currency symbols, big numeric values, replica, derivative, claim, property, wealth etc. We have an idea of some common words. Or we can simply collect some 10-15 high frequency words from the spam emails in the training data.

The naïve base classifier is a 3-step process, viz.

a. Computing the probability that the message is spam, knowing that a given word appears in this message;
b. Computing the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them);
c. Dealing with rare words.

As a natural rule **we do not consider the stop words like helping verbs, propositions etc. into consideration.** For the purpose of above three steps we make use of the Bayes theorem in an absolute manner and hence the name Naïve Bayes Classifier.

a. Computing the probability that the message is spam, knowing that a given word appears in this message:

Let's suppose the suspected message contains the word 'replica'. Most people who are used to receiving e-mail know that this message is likely to be spam, more precisely a proposal to sell counterfeit copies of well-known brands of watches (say). The spam detection software, however, does not "know" such facts; all it can do is compute probabilities. The formula used by the software to determine that is derived from Bayes' theorem:

$$\boldsymbol{P(S|W) = \frac{P(W|S).P(S)}{P(W|S).P(S) + P(W|H).P(H)}}$$

$P(S|W)$, is the probability that a message is a spam, knowing that the word 'replica' is in it;$P(S)$, is the overall probability that any given message is spam;$P(W|S)$, is the probability that the word 'replica' appears in spam messages;$P(H)$, is the overall probability that any given message is not spam (is "ham");$P(W|H)$, is the probability that the word 'replica' appears in ham messages.

The 'spamicity' of a word:

Recent statistics show that the current probability of any message being spam is 80% at the very least, which implies the following:

$$\boldsymbol{P(S) = 0.80 \; and \; P(H) = 0.2 : Prior \; Distribution/Probabilities}$$

However, most Bayesian spam detection software makes the assumption that there is no *a priori* reason for any incoming message to be spam rather than ham, and considers both cases to be equally likely.

$$\boldsymbol{P(S) = 0.50 \; and \; P(H) = 0.50 : Prior \; Distribution/Probabilities}$$

The filters that use this hypothesis are said to be "*not biased*", meaning that they have no prejudice regarding the incoming email. This assumption permits simplifying the general formula to the following:

$$P(S|W) = \frac{P(W|S)}{P(W|S) + P(W|H)}$$

This is functionally equivalent to asking: "*What percentage of occurrences of the word 'replica' appears in spam messages?*"This quantity is called "*spamicity*" (or "*spaminess*") of the word 'replica', and can be computed.

The number $P(W|S)$ used in this formula is approximated to the frequency of messages containing 'replica' in the messages identified as spam during the learning phase. Similarly, $P(W|H)$ is approximated to the frequency of messages containing 'replica' in the messages identified as ham during the learning phase.

b. Computing the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them):

Most Bayesian spam filtering algorithms are based on formulas that are strictly valid (from a probabilistic standpoint) only if the words present in the message are independent events. This condition is not generally satisfied (for example, in natural languages like English the probability of finding an adjective is affected by the probability of having a noun), but it is a useful idealization, especially since the statistical correlations between individual words are usually not known. On this basis, one can derive the following formula from Bayes' theorem:

$$\boldsymbol{p = \frac{p_1 p_2 \dots p_n}{p_1 p_2 \dots p_n + (1 - p_1)(1 - p_2) \dots (1 - p_n)}}$$

p is the probability that the suspect message is spam;p_1 is the probability $p(S|W_1)$ that it is a spam knowing it contains a first word (for example 'replica');p_2 is the probability $p(S|W_2)$ that it is a spam knowing it contains a second word (for example 'watches');p_n is the probability $p(S|W_n)$ that it is a spam knowing it contains an nth word (for example 'home').

This is the formula referenced by *Paul Graham* in his 2002 article. Spam filtering software based on this formula is sometimes referred to as a ***Naïve Bayes classifier.*** The result p is typically compared to a given threshold to decide whether the message is spam or not. If p is lower than the threshold, the message is considered as likely ham, otherwise it is considered as likely spam.

Case 1: Calculate the overall spamicity of the following emails and classify them as spam or non-spam. Assume that spam and non-spam emails are equally probable in nature.

Email 1:

Congratulations on winning the $ 100,000,000 in the lottery. To claim the prize, send your contact details to lucky@xyz.com.

Email 2:

Everything is going fine. I will not be coming for summer holidays. Take care of yourself.

	Word	P(W | S)	P(W | H)	P(W| S)+P(W | H)	p_i	1-p_i
1	Congratulations	0.8	0.2	1	0.80	0.20
1	winning	0.7	0.4	1.1	0.64	0.36
1	$	0.9	0.2	1.1	0.82	0.18
1	100000000	0.7	0.1	0.8	0.88	0.13
1	lottery	0.6	0.2	0.8	0.75	0.25
1	claim	0.6	0.3	0.9	0.67	0.33
1	prize	0.6	0.4	1	0.60	0.40
1	send	0.5	0.5	1	0.50	0.50
1	you	0.7	0.3	1	0.70	0.30
1	contact	0.5	0.5	1	0.50	0.50
1	details	0.6	0.6	1.2	0.50	0.50
2	Everything	0.2	0.7	0.9	0.22	0.78
2	going	0.2	0.7	0.9	0.22	0.78
2	fine	0.7	0.5	1.2	0.58	0.42
2	I	0.5	0.5	1	0.50	0.50
2	coming	0.5	0.5	1	0.50	0.50
2	summer	0.6	0.6	1.2	0.50	0.50
2	holidays	0.8	0.4	1.2	0.67	0.33
2	Take	0.7	0.6	1.3	0.54	0.46
2	care	0.2	0.2	0.4	0.50	0.50
2	Yourself	0.8	0.7	1.5	0.53	0.47

From the above table we conclude that:

For e-mail 1: p= 0.999784, which implies that it is close to 1, thus e-mail 1is a spam.

For e-mail 2: p= 0.233577 which is very small, thus e-mail 2 is a ham.

Email Spam Filtering using Logistic Regression

Prediction of Cancer due to Smoking using Logistic Regression

Given the data on a binary response variable telling us whether the cancer is present or not and a single binary independent variable telling whether the person smokes or not we want to predict the possibility of cancer due to smoking. In nutshell we want to know "how more likely is a person to have cancer if he/she smokes rather he/she doesn't". We are supposed to do the following:
A study was performed on lung cancer possibility due to smoking habits. Data on presence/absence of two attributes viz. lung cancer and smoking was collected for 25 individuals.
Case 3: Consider the dataset Smoking and Cancer.xlsx and perform the following objectives.

1. Build a logistic regression model for cancer possibility using smoking as an independent variable.

2. Test for the Significance of independent variable.

3. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given emails. Also calculate specificity and sensitivity of the model.

4. For each person obtain the probability of him/her having cancer and hence the predictionof cancer using the Logistic Classifier you have built.

5. Estimate the odds ratio and interpret it.

Step 1: Fitted logistic model:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X	.770	.823	.875	1	.350	2.160
	Constant	-.182	.606	.091	1	.763	.833

a. Variable(s) entered on step 1: X.

Fitted logistic regression model for cancer possibility using smoking as an independent variable:

$$\hat{\pi}_i = \frac{e^{(-0.182+0.770x)}}{1+e^{(-0.182+0.770x)}}$$

Step 2: Test for the Significance of independent variable

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Comparing the p-value (sig.) for smoking(X) with 0.05(p>0.05), we conclude that the variable X is not significant at 5% l.o.s.

Step 3: Confusion (Classification) Table

Classification Table[a]

Observed			Predicted		
			Lung Cancer		Percentage Correct
			0	1	
Step 1	Lung Cancer	0	6	5	54.5
		1	5	9	64.3
	Overall Percentage				60.0

a. The cut value is .500

- **Percentage of correct classification is = 60%**
- **Specificity** of the model is : **9/(9+5) =64.28%**, so we can infer that approx 65 out of 100 persons without lung cancer were correctly predicted to be negative for lung cancer.
- **Sensitivity** of the model is : **6/(6+5) = 54.545%**, that means approx 55 out of 100 persons known to have lung cancer were correctly predicted to have the lung cancer.

Step 4:

For each person obtain the probability of him/her having cancer and hence the predictionof cancer using the Logistic Classifier built above is in following table.

S.no	Predicted probability	Predicted Group
1	0.45455	0
2	0.45455	0
3	0.64286	1
4	0.64286	1
5	0.45455	0
6	0.45455	0
7	0.45455	0
8	0.45455	0
9	0.64286	1
10	0.64286	1
11	0.45455	0
12	0.45455	0
13	0.64286	1
14	0.64286	1
15	0.64286	1
16	0.45455	0
17	0.64286	1
18	0.64286	1
19	0.64286	1
20	0.45455	0
21	0.64286	1
22	0.64286	1
23	0.45455	0
24	0.64286	1
25	0.64286	1

Step 5: Odds Ratio for the variable X is: exp (β_1) =OR= 2.160 > 1, i.e there is associatrion or if a person smokes, he/she has a higher chance or is more likely to have Lung Cancer than the person who does not smoke or is not exposed to smoking.

Case 4: Skull Type Prediction using Logistic Regression

We are interested in predicting the type of skull of humans as one of two possible types I and II based on some five physical measures available related to the skulls.

Consider the dataset **Skull Type Prediction.xlsx** and perform the following objectives.

1. Build a logistic regression model for classifying a human skull as Type I/Type II using the given independent variables.
2. Test for the Significance individual independent variables.

3. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It's a Chi-Square Test).

4. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given skulls. Also calculate specificity and sensitivity of the model.

5. For each skull obtain the probability of it being Type I or Type II, and hence predict the skull Type using the Logistic Classifier you have built.

6. For a set of five physical measures given for a new skull in the dataset Skull Type Prediction – Validation Data.xlsx predict the skull type using the Logistic Classifier you have built.

Step1: Fitting logistic regression model:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X1	-.008	.018	.185	1	.668	.992
	X2	-.047	.033	2.039	1	.153	.954
	X3	-.007	.020	.113	1	.737	.993
	X4	-.006	.024	.054	1	.816	.994
	X5	.022	.020	1.245	1	.264	1.022
	Constant	1.149	2.826	.165	1	.684	3.155

a. Variable(s) entered on step 1: X1, X2, X3, X4, X5.

Fitted logistic regression model is:

$$\hat{\pi}_i = \frac{e^{(1.149-0.008X1-0.047X2-0.007X3-0.006X4+0.022X5)}}{1+e^{(1.149-0.008X1-0.047X2-0.007X3-0.006X4+0.022X5)}}$$

Step 2: Testing for the Significance individual independent variables:

H_0: $\beta_i = 0$ (i=1, 2, 3, 4, 5)

H_1: $\beta_i \neq 0$ for at least one i

Comparing the p-values (sig.) for XI (I=1, 2, 3, 4, 5) with 0.05, we conclude that all the independent variables are not significant, since p>0.05 for all variables.

Step 3: Testing for the overall Logistic Regression using **Hosmer and Lemeshow Test**

Contingency Table for Hosmer and Lemeshow Test

		skull_type = 0		skull_type = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2	1.841	0	.159	2
	2	2	1.774	0	.226	2
	3	2	1.446	0	.554	2
	4	0	1.374	2	.626	2
	5	2	1.226	0	.774	2
	6	1	1.073	1	.927	2
	7	0	.891	2	1.109	2
	8	1	.778	1	1.222	2
	9	1	.451	1	1.549	2
	10	0	.146	1	.854	1

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.601	8	.294

We conclude that since p-value for H-L test is 0.294> 0.05. Thus, fiited logistic regression model is a good fit for the given data.

Step 4:

Confusion (Classification) Table

Classification Table[a]

			Predicted		
			skull_type		Percentage Correct
	Observed		0	1	
Step 1	skull_type	0	9	2	81.8
		1	3	5	62.5
	Overall Percentage				73.7

a. The cut value is .500

- Percentage of correct classification in the given skulls is = 73.7%
- Specificity of the model is : 9/11= 81.8%,that means 81 out of 100 persons having skull type I were predicted to have the same skull type I
- Sensitivity of the model is: 5/8 = 62.5 %,that means 62 out of 100 persons having skull type II (observed) were predicted to have same skull type II

Step 5:

- ❖ Probability of skull type being Type I is: $11/19 = 0.58$
- ❖ Probability of skull type being Type II is: $8/19 = 0.42$
- ❖ Predictive probability of skull type being Type I is: $9/12 = 0.75$
- ❖ Predictive probability of skull type being Type II is: $2/7 = 0.29$

Prediction of skull Type using the Logistic Classifier we have built.

S.no	Predictive Probability	Predictive Group
1	0.58476	1
2	0.82867	1
3	0.07618	0
4	0.72021	1
5	0.08238	0
6	0.25299	0
7	0.32244	0
8	0.52374	1
9	0.3039	0
10	0.34643	0
11	0.6031	1
12	0.47311	0
13	0.61877	1
14	0.08545	0
15	0.30121	0
16	0.85355	1
17	0.4543	0
18	0.42803	0
19	0.14077	0

Step 6:

For a set of five physical measures, given below, for a new skull in the dataset Skull Type Prediction, predict the skull type using the Logistic Classifier we have built.

Substituing es

We have $\hat{ß}_0 = 1.149$, (X1=171, $\hat{ß}_1 = -0.008$),(X2=134, $\hat{ß}_2 = -0.047$),(X3=130, $\hat{ß}_3 = -.007$),(X4=69, $\hat{ß}_4 = -.006$) and (X5=130, $\hat{ß}_5 = .022$).

Fitted model: Now substituing above in model, we get

$$\hat{\pi}_i = \frac{e^{(1.149-0.008X_1-0.047X_2-0.007X_3-0.006X_4+0.022X_5)}}{1+e^{(1.149-0.008X_1-0.047X_2-0.007X_3-0.006X_4+0.022X_5)}}$$

We get $\hat{\pi}_i = 0.0068$, it is close to zero, thus we take Skull type to be of Type I.

Case 5: Sentiment Analysis using Logistic Regression – What makes a US Presidential Candidate Win?

What we are interested here in knowing that depending upon what and how a politician give speeches, his/her chances of winning the elections are affected. The idea here is similar to the email spam detection. The speech and more explicitly the content of the speech and it is delivery will have the information about the fact that the audience is convinced enough to vote for or against him/her. Sentiment Analysis is a discipline in itself; we are trying to understand the basics of to solve a particular problem. Commonly if politician is polite but passionate enough to serve the people, talks about development, remain optimist in his speech, talks about facts and figures related to government policies to explain his point to the audience is expected to win and vice-versa. But we want to examine what does data say?

The first aspect of the problem is to understand the data itself. I hope there is no confusion that we are going to use past data meaning by past win/loss statistics and the corresponding speeches. Clearly the response variable will indicate the win/loss information.

But what will be my independent variables? The independent variables will be the characteristics of the speech which may affect the win/loss which are commonly the following:

1. Proportion of words in the speech showing *Optimism*
2. Proportion of words in the speech showing *Pessimism*
3. Proportion of words in the speech showing the use of *Past*
4. Proportion of words in the speech showing the use of *Present*
5. Proportion of words in the speech showing the use of *Future*
6. Number of time he/she mentions his/her own party
7. Number of time he/she mentions his/her opposite parties

There are some more independent variables possible for which we need to understand the concept of big five personality traits which represent the personality traits of human which are the following:

A. Openness: *Curious, original, intellectual, creative and open to new ideas.*

B. Conscientiousness: *Organized, systematic, punctual, achievement oriented and dependable.*

C. Extraversion: *Outgoing, talkative, social and enjoys being in social situations.*

D. Agreeableness: *Affable, tolerant, sensitive, trusting, kind and warm.*

E. Neuroticism: *Anxious, irritable, temperamental and moody.*

Other that these big five personality traits the emotional content of the speech may also affect the win/loss. Thus we consider the following more independent variables.

8. Some measure indicating the content of speech showing *Openness*

9. Some measure indicating the content of speech showing *Conscientiousness*

10. Some measure indicating the content of speech showing *Extraversion*

11. Some measure indicating the content of speech showing *Agreeableness*

12. Some measure indicating the content of speech showing *Neuroticism*

13. Some measure indicating the content of speech showing *emotionality*

Once we get this data, task is all with the statistical analyst to make an efficient model with good predictive power.

Case 5: Consider the **US Presidential Data.xlsx** and perform the following objectives:

1. Build a logistic regression model for classifying win/loss using the given independent variables.

2. Test for the Significance individual independent variables.

3. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It's a Chi-Square Test).

4. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given skulls. Also calculate specificity and sensitivity of the model.

5. For each speech obtain the probability of winning, and hence predict the win/loss status using the Logistic Classifier you have built.

Step 1: Fitting Logistic Regression Model:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Optimism(X1)	-3.567	2.090	2.912	1	.088	.028
	Pessimism(X2)	-28.451	2.951	92.952	1	.000	.000
	PastUsed(X3)	2.080	.763	7.439	1	.006	8.002
	FutureUsed(X4)	4.138	.725	32.559	1	.000	62.668
	OwnPartyCount(X5)	.008	.006	1.756	1	.185	1.008
	OppPartyCount(X6)	.016	.013	1.466	1	.226	1.016
	NumericContent(X7)	311.989	53.759	33.680	1	.000	3.128E+135
	Extra(X8)	-.377	.082	20.967	1	.000	.686
	Emoti(X9)	.212	.130	2.669	1	.102	1.237
	Agree(X10)	-.583	.188	9.649	1	.002	.558
	Consc(X11)	-.481	.118	16.770	1	.000	.618
	Openn(X12)	.828	.107	60.040	1	.000	2.288
	Constant	.936	.768	1.483	1	.223	2.549

a. Variable(s) entered on step 1: Optimism, Pessimism, PastUsed, FutureUsed, OwnPartyCount, OppPartyCount, NumericContent, Extra, Emoti, Agree, Consc, Openn.

Fitted logistic regression model is:

$$\hat{\pi}_i = \frac{e^{(0.936-3.567X1-28.451X2+2.080X3+4.138X4+0.008X6+0.016X7+311.989X8-0.377X9+0.212X10-0.583X11-0.481X12+0.828X13)}}{1+e^{(0.936-3.567X1-28.451X2+2.080X3+4.138X4+0.008X6+0.016X7+311.989X8-0.377X9+0.212X10-0.583X11-0.481X12+0.828X13)}}$$

Step 2: Testing for the Significance individual independent variables

H_0: $\beta_i = 0$ (i=0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)

H_1: $\beta_i \neq 0$ for at least one i

We conclude for variables Pessimism(X_2), Pastused(X_3), FutureUsed(X_4), NumericContent(X_7), Extra(X_8), Agree(X_{10}), Consc(X_{11}) and Openn(X_{12}) are significant as their p-values are less than 0.05.

Step 3: Test for the overall Logistic Regression using Hosmer and Lemeshow Test.

Contingency Table for Hosmer and Lemeshow Test

		Win/Loss = 0		Win/Loss = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	127	122.009	25	29.991	152
	2	113	102.257	39	49.743	152
	3	91	87.973	61	64.027	152
	4	57	74.678	95	77.322	152
	5	62	62.229	90	89.771	152
	6	55	49.828	97	102.172	152
	7	25	39.285	127	112.715	152
	8	28	29.436	124	122.564	152
	9	16	19.326	136	132.674	152
	10	21	7.978	135	148.022	156

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	43.905	8	.000

From the H- L test, we conclude that p-value (Sig.) is 0.000 < 0.05 (l.o.s), thus fitted logistic regression model does not give a good fit to the given data.

Step 4:

Confusion (Classification) Table

Classification Table[a]

			Predicted		
			Win/Loss		Percentage Correct
	Observed		0	1	
Step 1	Win/Loss	0	354	241	59.5
		1	157	772	83.1
	Overall Percentage				73.9

a. The cut value is .500

- Percentage of correct classification is = (354 + 772)/(354 + 241+157+772) = 73.88%
- Specificity of the model is: 354 354+157 = 69.28%, so we can infer that 69 out of 100 candidates who lost were correctly predicted to lose.
- Sensitivity of the model is : 772 772+241 = 76.21%, that means 76 out of 100 candidates known to have won elections were correctly predicted to win.

Discriminant Analysis

Discriminant Analysis is a classification technique to classify an observation (univariate or multivariate) into one of several possible classes by means of "some" optimal way to separate different populations. Some real life examples of classification problem are loan classification – high risk, medium risk and low risk, warning systems for financial crisis, medical diagnostics – critical and non-critical patients.

Suppose we have $x_1, \dots, x_n$ as n multivariate observations and let $\mathcal{X}$ be the measurement space of all the multivariate observations. Further, suppose that each of the observations fall into one (exactly one) of the J classes denoted as $C = \{1, \dots, J\}$: Set of classes.

In discriminant analysis the basic aim is develop a systematic way of predicting the class membership of a multivariate observation by means of some optimal classification rule.

Definition: A classifier or a classification rule is a function $d(x)$ defined on $\mathcal{X}$ such that for every $x \in \mathcal{X}$, $d(x)$ is equal to one of the numbers $1, \dots J$.

Alternate way to look at the classifier is that it induces a partition of the entire measurement space $\mathcal{X}\{A_1, \dots, A_J\}$ such that

$$A_j = \{x : d(x) = j\}\,; j = 1, \dots, J$$

Let us concentrate on the binary classification problem. Suppose there are two populations π_1 and π_2 and an arbitrary multivariate observation comes from either of the two.

Aim: Let x_1 and x_2 be observations from π_1 andπ_2, the aim is to find some function say g such that $g(x_1)$ and $g(x_2)$ look as different as possible then g is the desired discriminant function to discriminate between π_1 andπ_2. The aim is to find some "optimal" discriminant function. One such optimal rule is given by fisher linear discriminant function.

1.1.1 Fisher Linear Discriminant Analysis

Under the same setup assume that, $X \mid \pi_1 \sim (\mu_1, \Sigma)$ where μ_1 is the mean vector of the 1st population and Σ is the covariance matrix for the 1st population, and $X \mid \pi_2 \sim (\mu_2, \Sigma)$ where μ_2 is the mean vector of the 2nd population and Σ is the common covariance matrix for both the populations.

Further change π_1 and π_2 into two univariate populations by changing X to some $l'X$ by means of "some" l.

$$X \mid \pi_1 \sim (\mu_1, \Sigma) \Rightarrow\ l'X \mid \pi_1 \sim (l'\mu_1, l'\Sigma l)$$

$$X \mid \pi_2 \sim (\mu_2, \Sigma) \Rightarrow\ l'X \mid \pi_2 \sim (l'\mu_2, l'\Sigma l)$$

Discrimination:

We are interested in finding or choosing $l$such that the separation between the two univariate populations is maximum, i.e. maximization of statistical distance between π_1 and π_2 with respect to l.

A measure of statistical distance between the two populations is given by,

$$\frac{\left(l'\mu_1 - l'\mu_2\right)^2}{l'\Sigma l} = \frac{(l'(\mu_1 - \mu_2))^{\,2}}{l'\Sigma l}$$

We want to obtain,

$$l = \arg\max_{l} \frac{(l'(\mu_1 - \mu_2))^{\,2}}{l'\Sigma l}$$

Assuming that Σ is positive definite and defining $a' = l'\Sigma^{1/2}$. We have,

$$\frac{(a'\Sigma^{-1/2}(\mu_1 - \mu_2))^{\,2}}{a'a} \ldots (*)$$

Using Cauchy Schetwartz Inequality,

$$\frac{(a'\Sigma^{-1/2}(\mu_1 - \mu_2))^{\,2}}{a'a} \leq \frac{\left(a'a\right)\left((\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\right)}{a'a}$$

$$= (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\text{: Mahalanobis Distance}$$

Hence we have the distance between two populations is always less than or equal to the Mahalanobis Distance. $\frac{(l'(\mu_1-\mu_2))^{\,2}}{l'\Sigma l}$is maximum when,

$$a' = (\mu_1 - \mu_2)'\Sigma^{-1/2} \Rightarrow l' = (\mu_1 - \mu_2)'\Sigma^{-1} : \text{optimal } l$$

Thus we have the optimal l which provides the maximum separation (discrimination) between the two populations.

The quantity $l'X = (\mu_1 - \mu_2)'\Sigma^{-1}X$ is called the **Fisher Linear Discriminant Function (LDF),** which is an optimal separation between the two populations.

Now then next question of interest is that based on Fisher LDF how do we classify an arbitrary observation into one of the two possible classes?

Classification:

Realize that,

$$E\left[l'X \,|\pi_i\right] = E[(\mu_1 - \mu_2)'\Sigma^{-1}X \,|\pi_i] = (\mu_1 - \mu_2)'\Sigma^{-1}\mu_i \coloneqq m_i \; (say)$$

Note that

$$m_1 - m_2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \geq 0 \text{ as } \Sigma \text{ is pd.}$$

$$\Rightarrow m_1 \geq m_2$$

As a rule of classification for any new observation x_0 calculate $y_0 = (\mu_1 - \mu_2)'\Sigma^{-1}x_0 =$ Fisher LDF of x_0, and assign x_0 to π_1 if y_0 is closer to m_1 than m_2 otherwise to π_2 which is an intuitive logical rule as we are trying to see if the FLDF is close to expectation of FLDF under π_1 or π_2. Thus we have,

$\pi_2$region $\pi_1$region

$$m_2 \quad \frac{m_1 + m_2}{2} \quad m_1$$

We can finally write that assign x_0 to π_1 if

$$y_0 = (\mu_1 - \mu_2)'\Sigma^{-1}x_0 > \frac{m_1 + m_2}{2}$$

$$y_0 > \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$$

and assign x_0 to π_2 otherwise.

Usually for practical problems the values of μ_1, μ_2 and Σ are unknown, which we replace by their estimates. For estimation we need samples from both the populations, i.e. for some data points we need to have the classes being already assigned, i.e. we have data of the following form:

$$(x_1, j_1), \ldots, (x_n, j_n)$$

Where $j_i = 1$ or 2 representing the class or the population.

$$\widehat{\mu_1} = \frac{1}{\#[i : j_i = 1]} \sum_{i:j_i=1} x_i \text{ and } \widehat{\mu_2} = \frac{1}{\#[i : j_i = 2]} \sum_{i:j_i=2} x_i$$

and$\hat{\Sigma}$ is calculated as the pooled sample variance of the observations from both the samples. So finally the classifier in its executable form can be written as:

$$\text{if}(\widehat{\mu_1} - \widehat{\mu_2})'\hat{\Sigma}^{-1}x_0 > \frac{1}{2}(\widehat{\mu_1} - \widehat{\mu_2})'\hat{\Sigma}^{-1}(\widehat{\mu_1} + \widehat{\mu_2}) \text{ then class is } \pi_1 \text{ otherwise } \pi_2.$$

The FLDF ofx_0, i.e. $(\mu_1 - \mu_2)'\Sigma^{-1}x_0$with population characteristics being replaced by their sample counterparts is called as the sample FLDF.

So far we have discussed the theoretical basis and intuitive idea behind FLDA. Let us now concentrate on how we carry it out in SPSS.

Assumptions of FLDA:

The assumptions of discriminant analysis are the same as those for MANOVA. The analysis is quite sensitive to outliers and the *size of the smallest group must be larger than the number of predictor variables*.

- **Multivariate normality:** Independent variables are normal for each level of the grouping variable.
- **Homogeneity of variance/covariance (homoscedasticity):** Variances among group variables are the same across levels of predictors. This can be tested with Box's M statistic.
- **Multicollinearity:** Predictive power can decrease with an increased correlation between predictor variables.
- **Independence:** Participants are assumed to be randomly sampled, and a participant's score on one variable is assumed to be independent of scores on that variable for all other participants.

It has been suggested that discriminant analysis is relatively robust to slight violations of these assumptions, and it has also been shown that discriminant analysis may still be reliable when using dichotomous variables (where multivariate normality is often violated)

Some Discussion on the Tests and Routines to be used:

1. Wilk's Lambda

Wilk's Lambda is a test for equality of group means and is used to test which independent variable contributes significantly to the discriminant function, i.e. which variable contributes significantly while discriminating between the 2 groups 0 & 1.

H_0: All the group means are statistically not significantly different between the 2 groups for a particular measure.
H_1: All the group means are statistically significantly different between the 2 groups for a particular measure.

The value of Wilk's Lambda varies between 0 and 1. The smaller its value, the more the corresponding variable contributes to the discriminant function.

2. Box's M test:

Box's M test tests for the homogeneity of variance-covariance matrices of the 2 groups.

Ho: Covariance matrices of the 2 groups do not differ significantly.
$H1$: Covariance matrices of the 2 groups differ significantly.

This is a very powerful test, so when the sample is large then even small differences are considered significant. Thus, in order to be lenient, we check for values of Log Determinants. If these values for the 2 groups are fairly close, we accept H_o and conclude that the 2 covariance matrices are not significantly different.

3. Summary of Canonical Discriminant Functions

a. Eigenvalue and Canonical correlation:

Eigenvalue represents the ratio of the between-group sum of squares to the within-group sum of squares of the discriminant score. It indicates the *relative discriminating power of the discriminant function*, i.e. how well the discriminating function discriminates between the 2 groups.

Canonical Correlation of discriminant function is the correlation of that function with discriminant scores.Since there are only 2 groups so only one discriminant function is generated which accounts for 100% of the explained variance.If Canonical Correlation is close to 1, it implies nearly all the variation in the discriminant scores can be attributed to the group differences. *Squared canonical correlation is the percentage of variation in the dependent, discriminated by the set of independents in discriminant analysis*.

b. Standardized Canonical Discriminant Function Coefficients:

Standardized Canonical Discriminant Function Coefficients allows us to see the extent to which each of the predictors contribute to the ability of the discriminant function. It rescales the variables to unit standard deviation. If a coefficient lies in the neighborhood of 1 or -1, then it is a good explanator& if it lies in the neighborhood of 0 or 0.5, then it gives a moderate explanation. SPSS generates Unstandardized Canonical Discriminant Function Coefficients as well which gives the same information but as they are not standardized the same rule of interpretation is not applicable.

c. Functions at Group Centroids:

These are the estimated expected values of the FLDF for different groups. In machine learning terminology centroid is nothing but the mean. So, these are our m_1 and m_2 only.

4. Classification Statistics:

a. Classification Function Coefficients:

Recall that we need to calculate the optimum $l' = (\widehat{\mu_1} - \widehat{\mu_2})'\hat{\Sigma}^{-1}$. Hence the difference between the values of classification function coefficients for two different groups will give us the entries of l' which we can use to multiply with the values of a new observation to calculate sample FLDF and then compare it with $\frac{m_1+m_2}{2}$ to classify it into one of the two groups.

b. Classification Results:

SPSS generates two classification tables which are as follows:

(i) Original: which tells the fitting strength of the LDF or how well it performs for the given data?
(ii) Cross-Validated: which tells the predictive strength of the LDF or how well it performs for the new data?

Skull Type Prediction using Discriminant Analysis

We are interested in predicting the type of skull of humans as one of two possible types I and II based on some five physical measures available related to the skulls.

Case 6: Consider the dataset **Skull Type Prediction.xlsx** and perform the following objectives.

1. Test for normality of all five physical measures.
2. Test for equality of the covariance matrices of different groups using Box's M Test.
3. Test if the group means are statistically significantly different between the 2 groups for all the physical measures and point out which measures contribute to the discriminant function significantly.
4. Test for overall significance of the discriminant function.
5. Obtain the standardized and unstandardized canonical discriminant function coefficients.
6. Obtain the *values of discriminant function at the group centroids* and *classification function coefficients.*
7. Obtain the entries of the vector l, calculate the discriminant functions for the given skulls and hence obtain the predicted skull type for the given skulls.
8. Construct the classification tables for fitting strength and predictive strength of the model. Also calculate sensitivity and specificity for both fitting strength and predictive strengths.
9. For a set of five physical measures given for a new skull in the dataset **Skull Type Prediction – Validation Data.xlsx** predict the skull type using the Logistic Classifier you have built.

Step 1: Testing for Normality

Tests of Normality

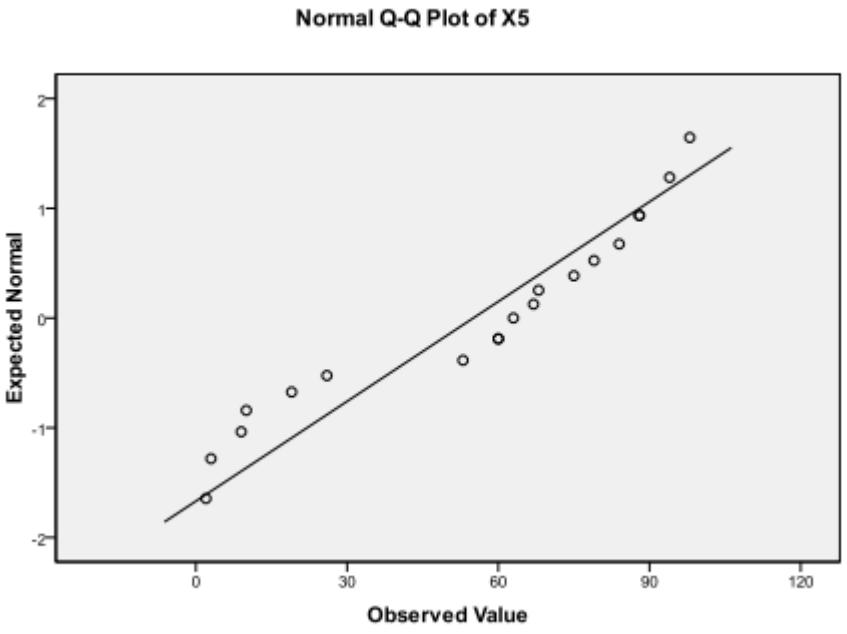
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
X1	.174	19	.133	.909	19	.071
X2	.112	19	.200*	.966	19	.702
X3	.131	19	.200*	.947	19	.344
X4	.127	19	.200*	.978	19	.913
X5	.191	19	.066	.890	19	.032

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

We conclude that variables X1, X2, X3 and X4 are normal both by K-S Test and S-W test as the p-values > 0.05 for both tests. But variable X5 is normal by K-S test but not by S-W test.

Consider Q-Q plot of the variable X5



Since points move away from the reference line or line of fit on both ends. Thus, variable X5 is not normal.

Step 2: Testing for equality of the covariance matrices of different groups using Box's M Test.

H_o: Covariance matrices of the groups do not differ significantly

H_1: Covariance matrices of the groups differ significantly

Test Results

Box's M		21.637
F	Approx.	.947
	df1	15
	df2	908.297
	Sig.	.511

Tests null hypothesis of equal population covariance matrices.

Since p-value for Box's M test is 0.511 >0.05(l.o.s). Thus, the null hypothesis is accepted and thus, the variance-covariance matrices of groupsdo not differ significantly.

This is a very powerful test, so when the sample is large then even small differences are considered significant. Thus, in order to be lenient, we check for values of Log Determinants.

Log Determinants

skull_type	Rank	Log Determinant
0	5	32.677
1	5	29.947
Pooled within-groups	5	32.825

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

We conclude that values of log determinants of skull type 1 and skull type 2 do have some difference, thus covariance matrices of the groups differ significantly.

Step 3: Testing if the group means are statistically significantly different between the 2 groups for all the physical measures and point out which measures contribute to the discriminant function significantly.

H_o: All the group means are statistically not significantly different between the 2 groups for a particular measure.

H_1: All the group means are statistically significantly different between the 2 groups for a particular measure

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
x1	.986	.242	1	17	.629
x2	.859	2.794	1	17	.113
x3	1.000	.005	1	17	.943
x4	.989	.184	1	17	.674
x5	.888	2.134	1	17	.162

Since, we know that value of Wilk's Lambda varies between 0 and 1. The smaller its value, the more the corresponding variable contributes to the discriminant function.Thus, we can coclude from table that $X3$ has negligible contribution to discriminant function, whereas, $X2$ has the maximum contribution to the discriminant function

Step 4: Testing for overall significance of the discriminant function

H_o: Discriminant function is not significant

H_1: Discriminant function is significant.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.776	3.678	5	.597

Since p-value (Sig.)>0.05(l.o.s), we fail to reject null hypothesis i.e. We conclude that overall discriminant function is not significant. Since wilk's lambda is close to 1 , so the overall discriminant analysis does not contributes significantly as 77.6% of the variability in the discriminate scores has not been explained by the differences among groups.

Step 5: Obtain the standardized and unstandardized canonical discriminant function coefficients.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
X1	.239
X2	.840
X3	.154
X4	.176
X5	-.592

Standardized canonical discriminant functioncoefficients give the idea which of the variables contributes to the ability of the discriminant function. It can be seen that standardized coefficient of X2 is close to 1, thus, it is a good explanator, while the other standardized values are near about 0.5 or less than that and hence, are considered to be moderate explanatory variables.

Coefficients

	Function
	1
X1	.007
X2	.042
X3	.005
X4	.007
X5	-.019
(Constant)	-1.559

Unstandardized coefficients

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.289[a]	100.0	100.0	.473

a. First 1 canonical discriminant functions were used in the analysis.

Canonical correlation is 0.473(square of this value is 0.223729) which suggests that 22.37% of the variation is explained by the discriminant function. The discriminant function is not appropriate to discriminate the two groups. Also, Eigen value is given as 0.289, which is very much on the lower side implying between group sum of squares is very low compared to the within group sum of squares. Thus, the discriminant function is not appropriate in discriminating the two groups.

Step 6: Obtain the values of discriminant function at the group centroids and classification function coefficients.

Functions at Group Centroids

Skull_Type	Function
	1
.00	.433
1.00	-.596

Unstandardized canonical discriminant functions evaluated at group means

Values of group centroids are $m_1 = .433$ and $m_2 = -.596$

Classification Function Coefficients

	Skull_Type	
	.00	1.00
X1	.100	.093
X2	.233	.189
X3	.055	.050
X4	.139	.132
X5	.062	.081
(Constant)	-14.614	-13.093

Classification function coefficients are:

$l' = (0.007 \;\; 0.044 \;\; 0.005 \;\; 0.007 \;\; -0.019)$

Step 7: Obtain the entries of the vector l, calculate the discriminant functions for the given skulls and hence obtain the predicted skull type for the given skulls.

Discriminant function for given skulls is:

$$l'X = 0.007\,X1 + 0.044\,X2 + 0.005\,X3 + 0.007\,X4 - 0.019\,X5$$

Predicted skull type for the given skulls are in following table:

skull_type	X1	X2	X3	X4	X5	Predicted	Discriminant
1	13	29	82	24	60	1	-0.77244
0	79	1	1	56	63	1	-1.73441
0	5	77	47	45	26	0	1.81815
1	100	16	80	60	98	1	-1.1844
0	55	65	3	20	2	0	1.69458
0	91	55	20	59	68	0	0.66052
1	47	31	31	52	19	0	0.25397
1	17	43	61	45	79	1	-0.45431
1	30	54	11	83	60	0	0.48143
0	45	17	63	79	10	0	0.16713
0	1	40	97	51	94	1	-0.74671
1	69	44	40	47	84	1	-0.24014
1	95	19	33	2	53	1	-0.91126
0	83	47	75	68	9	0	1.69569
0	78	57	86	19	88	0	0.32209
1	94	1	50	44	88	1	-1.94052
0	7	54	15	23	67	1	-0.22034
0	77	45	34	32	75	1	-0.11529
0	30	37	81	92	3	0	1.22625

Step 8: Construct the classification tables for fitting strength and predictive strength of the model. Also calculate sensitivity and specificity for both fitting strength and predictive strengths.

Classification Results[b,c]

		Skull_Type	Predicted Group Membership .00	Predicted Group Membership 1.00	Total
Original	Count	.00	7	4	11
		1.00	2	6	8
	%	.00	63.6	36.4	100.0
		1.00	25.0	75.0	100.0
Cross-validated[a]	Count	.00	5	6	11
		1.00	5	3	8
	%	.00	45.5	54.5	100.0
		1.00	62.5	37.5	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 68.4% of original grouped cases correctly classified.

c. 42.1% of cross-validated grouped cases correctly classified.

We conclude from table that discriminant analysis predicts the correct skull type in 68.4% of the cases, whereas skull type is predicted in 42.1% of the cases correctly when cross validation is done. Hence, discriminant analysis is moderate in differentiating between the skull types.

For fitting strengths: (based on the original data)

- **Sensitivity** = 6/(6+2) = 0.75 = 75%
- **Specificity** = 7/(7+4) = 0.6364 = 63.64%

For predictive strengths :(Based on cross validated data)

- **Sensitivity =** 3/(3+5)= 0.375 = 37.5%
- **Specificity =** 5/(5+6)= 0.4546 = 45.46%

Step 9: For a set of five physical measures given for a new skull in the dataset Skull Type **Prediction –Validation Data.xlsx** predict the skull type using the Logistic Classifier you have built.

We have: X1 = 171, X2 = 134, X3 = 130, X4 = 69 and X5 = 130

l'X = 0.007*171+0.044*134+0.005*130+0.007*69-0.019*130 = 5.756 > $(m1+m2)/2$

(since m2<m1)

Thus, predicted skull type is Type I.

Case 7: Comparative Study of Binary Logistic Regression and Binary Discriminant Analysis:

The Skull Type Prediction Problem has already been solved using Logistic Regression and now you will be solving the same problem using a different technique Discriminant Analysis. Compare the two methodologies for Skull Type Prediction Problem on following grounds:

1. Classification of individual skull.

skull_type	Predicted(Binary Logistic)	Predicted(Discriminant)
1	1	1
0	1	1
0	0	0
1	1	1
0	0	0
0	0	0
1	0	0
1	1	1
1	0	0
0	0	0
0	1	1
1	**0**	**1**
1	1	1
0	0	0
0	0	0
1	1	1
0	**0**	**1**
0	**0**	**1**
0	0	0

We can conclude from the above table that 3 bold cases in which the predicted groups are different by obtaining from both the methods. On comparing from the original skull type data in one case the model fit by discriminant analysis proves to be better as it gives the correct predicted value whereas in the other two cases model fitted by binary logistic proves to be better.

2. Confusion Matrix, Percentage of Correct Classification

2. **Confusion Table[a] for Classification using Binary Logistic**

			Predicted		
			Skull_Type		Percentage
	Observed		.00	1.00	Correct
Step 1	Skull_Type	.00	9	2	81.8
		1.00	3	5	62.5
	Overall Percentage				73.7

a. The cut value is .500

Confusion Table for Classification using Discriminant Analysis

		Skull_Type	Predicted Group Membership		
			.00	1.00	Total
Original	Count	.00	7	4	11
		1.00	2	6	8
	%	.00	63.6	36.4	100.0
		1.00	25.0	75.0	100.0

On comparing percentages of skull types predicted correctly using Binary Logistic is 73.7% and that using Discriminant Analysis is 68.4%. Thus, Binary Logistic can be considered to be a better discriminating method of the two.

Also, if we compare the diagonal entries i.e. the correctly classified ones (or the non diagonal elements i.e. the misclassified ones) .Clearly in one case $9 > 7$, hence the binary logistic model is a good fit for skull type I, whereas $5 < 6$ hence the discriminant analysis method fits better model for skull type II.

3. Sensitivity and Specificity

	Binary Logistic	Discriminant Analysis
Sensitivity	62.5%	75%
Specificity	81.82%	63.64%

We conclude from above table that as we kniw more the sensitivity, better the model. So, the binary logistic model fit is better in terms of sensitivity as the model fit by binary logistic method is more sensitive,

Also, specificity of the model fitted by discriminant analysis is more, so it is a better model on the basis of specificity. The model obtained by discriminant analysis is more specific.

4. Performance on the Validation Data.

On comparing the results obtained in both the methods, we conclude that the validated data predicts the skull type to be Type I. Thus, on the basis of validation data, both the method seems to be equally efficient in predicting the skull type.

Multiclass Classification

We have already discussed that a classification problem is said to be multiclass classification problem if the response is has more than two possible classes. We will be studying only Multiclass Logistic Regression as a multiclass classification technique though the other two techniques viz. Naïve Bayes' Classifier and Discriminant Analysis as well have their multivariate extensions

Logistic Regression can be used to solve a multiclass classification problem in following two ways:

- By means of decomposing the multiclass classification problem into several binary classification problems.
- By means of using multinomial probability distribution.

Decomposing the multiclass classification problem into several binary classification problems:

Suppose the response variable has three class viz. 1, 2 and 3 then the response is defined as follows:

$$y_i = \begin{cases} 1 & \text{with prob } \pi_{1i} \\ 2 & \text{with prob} \pi_{2i} \\ 3 & \text{with prob} \pi_{3i} \end{cases}$$

We can define three binary variables using y_i as follows:

$$y_{1i} = \begin{cases} 1 & \text{if } y_i = 1 \\ 0 & \text{otherwise} \end{cases}, \quad y_{2i} = \begin{cases} 1 & \text{if } y_i = 2 \\ 0 & \text{otherwise} \end{cases} \text{ and } \quad y_{3i} = \begin{cases} 1 & \text{if } y_i = 3 \\ 0 & \text{otherwise} \end{cases}$$

Clearly we have three binary classification problems which model $P[y_{1i} = 1] = \pi_{1i}$, $P[y_{2i} = 1] = \pi_{2i}$and $P[y_{3i} = 1] = \pi_{3i}$ respectively.

Using Binary Logistic Regression we can obtain $\widehat{\pi_{1i}}, \widehat{\pi_{2i}}$ and $\widehat{\pi_{3i}}$ and can classify y_{1i}, y_{2i} and y_{3i} but our goal is to classify y_i the 3-class variable. We define the rule for multiclass classification as follows:

$$\widehat{y_i} = \underbrace{\text{argmax}}_{k} \widehat{\pi_{ki}}$$

Where $k = 1,2,3$ (# of classes) and $i = 1, \dots, n$ (# of observations).

Multinomial Distribution for Multiclass Logistic Regression Problem:

Suppose the response variable has three class viz. 1, 2 and 3 then the response is defined as follows:

$$y_i = \begin{cases} 1 & \text{with prob } \pi_{1i} \\ 2 & \text{with prob} \pi_{2i} \\ 3 & \text{with prob} \pi_{3i} \end{cases}$$

We can use the multinomial probability distribution to obtain the probability mass functions of y_is and hence the likelihood functions of the sample observations. We can define appropriate links for different probabilities with the predictor variables. Then the likelihood function can be maximized using the IRLS technique and we can proceed further in a similar manner. SPSS does all of it in a built-in routine.

Case 8: Consider the dataset **Flower Species.xlsx** dataset which has data on Sepal Length, Sepal Width, Petal Length and Petal Width and Species Type for 150 different flowers and perform the following objectives.

1. Decompose the multiclass (3-class) classification problem into three Binary Classification Problems and perform the following for each problem:
   a. Test for the Significance individual independent variables.
   b. Construct the Confusion (Classification) Table and report the percentage of correct classification for the given emails. Also calculate specificity and sensitivity of the model.
   c. For each flower obtain the predicted probability and hence the predicted class using the Logistic Classifier you have built.
2. Obtain the multiclass predicted flower species for the original problem using the three sub problems.
3. Construct the classification matrix and report the percentage of correct classification.

Define a new variable x1 which takes the value 1 when the flower species is setosa or 0 otherwise

Taking setosa as 1, versicolor and virginica as 0:

H_o: $\beta_i = 0$ (i=1, 2, 3, 4)

H_1: $\beta_i \neq 0$ for atleast one of the i

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	Sepal_Length	8.666	14840.420	.000	1	1.000
	Sepal_Width	6.637	6922.781	.000	1	.999
	Petal_Length	-15.119	12366.721	.000	1	.999
	Petal_Width	-16.272	17915.189	.000	1	.999
	Constant	-13.616	51859.146	.000	1	1.000

The p-values for all the variables Sepal Length, Sepal Width, Petal Length and Petal Width are > 0.05 (l.o.s). We fail to reject our null hypothesis and thus, we conclude none of the variables are significant.

a. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It's a Chi-Square Test).

Since we cannot reject all the variables, now testing:

H_o: $\beta = 0$

H_1: $\beta \neq 0$

Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	.000	8	1.000

Since p-value for H-L test comes out to be 1, which is >0.05(l.o.s). Thus, we fail to reject our null hypothesis, and thus, the fitted model is a good fit.

Classification Table[a]

Observed			Predicted		
			m1		Percentage Correct
			.00	1.00	
Step 1	m1	.00	100	0	100.0
		1.00	0	50	100.0
		Overall Percentage			100.0

a. The cut value is .500

Percentage of correct classification for given e-mails is 100%. Hence, the fitted model is the best possible model.

- Sensitivity= 50/(50+0) = 1
- Specificity = 100/(100+0)=1

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	x1	Pre_prob	Pre_grp
1	5.1	3.5	1.4	0.2	setosa	1	1	1
2	4.9	3	1.4	0.2	setosa	1	1	1
3	4.7	3.2	1.3	0.2	setosa	1	1	1
4	4.6	3.1	1.5	0.2	setosa	1	1	1
5	5	3.6	1.4	0.2	setosa	1	1	1
6	5.4	3.9	1.7	0.4	setosa	1	1	1
7	4.6	3.4	1.4	0.3	setosa	1	1	1
8	5	3.4	1.5	0.2	setosa	1	1	1
9	4.4	2.9	1.4	0.2	setosa	1	1	1

10	4.9	3.1	1.5	0.1	setosa	1	1	1
11	5.4	3.7	1.5	0.2	setosa	1	1	1
12	4.8	3.4	1.6	0.2	setosa	1	1	1
13	4.8	3	1.4	0.1	setosa	1	1	1
14	4.3	3	1.1	0.1	setosa	1	1	1
15	5.8	4	1.2	0.2	setosa	1	1	1
16	5.7	4.4	1.5	0.4	setosa	1	1	1
17	5.4	3.9	1.3	0.4	setosa	1	1	1
18	5.1	3.5	1.4	0.3	setosa	1	1	1
19	5.7	3.8	1.7	0.3	setosa	1	1	1
20	5.1	3.8	1.5	0.3	setosa	1	1	1
21	5.4	3.4	1.7	0.2	setosa	1	1	1
22	5.1	3.7	1.5	0.4	setosa	1	1	1
23	4.6	3.6	1	0.2	setosa	1	1	1
24	5.1	3.3	1.7	0.5	setosa	1	1	1
25	4.8	3.4	1.9	0.2	setosa	1	1	1
26	5	3	1.6	0.2	setosa	1	1	1
27	5	3.4	1.6	0.4	setosa	1	1	1
28	5.2	3.5	1.5	0.2	setosa	1	1	1
29	5.2	3.4	1.4	0.2	setosa	1	1	1
30	4.7	3.2	1.6	0.2	setosa	1	1	1
31	4.8	3.1	1.6	0.2	setosa	1	1	1
32	5.4	3.4	1.5	0.4	setosa	1	1	1
33	5.2	4.1	1.5	0.1	setosa	1	1	1
34	5.5	4.2	1.4	0.2	setosa	1	1	1
35	4.9	3.1	1.5	0.2	setosa	1	1	1
36	5	3.2	1.2	0.2	setosa	1	1	1
37	5.5	3.5	1.3	0.2	setosa	1	1	1
38	4.9	3.6	1.4	0.1	setosa	1	1	1
39	4.4	3	1.3	0.2	setosa	1	1	1
40	5.1	3.4	1.5	0.2	setosa	1	1	1
41	5	3.5	1.3	0.3	setosa	1	1	1
42	4.5	2.3	1.3	0.3	setosa	1	1	1
43	4.4	3.2	1.3	0.2	setosa	1	1	1
44	5	3.5	1.6	0.6	setosa	1	1	1
45	5.1	3.8	1.9	0.4	setosa	1	1	1
46	4.8	3	1.4	0.3	setosa	1	1	1
47	5.1	3.8	1.6	0.2	setosa	1	1	1
48	4.6	3.2	1.4	0.2	setosa	1	1	1
49	5.3	3.7	1.5	0.2	setosa	1	1	1
50	5	3.3	1.4	0.2	setosa	1	1	1
51	7	3.2	4.7	1.4	versicolor	0	0	0
52	6.4	3.2	4.5	1.5	versicolor	0	0	0
53	6.9	3.1	4.9	1.5	versicolor	0	0	0
54	5.5	2.3	4	1.3	versicolor	0	0	0
55	6.5	2.8	4.6	1.5	versicolor	0	0	0
56	5.7	2.8	4.5	1.3	versicolor	0	0	0
57	6.3	3.3	4.7	1.6	versicolor	0	0	0

58	4.9	2.4	3.3	1	versicolor	0	0	0
59	6.6	2.9	4.6	1.3	versicolor	0	0	0
60	5.2	2.7	3.9	1.4	versicolor	0	0	0
61	5	2	3.5	1	versicolor	0	0	0
62	5.9	3	4.2	1.5	versicolor	0	0	0
63	6	2.2	4	1	versicolor	0	0	0
64	6.1	2.9	4.7	1.4	versicolor	0	0	0
65	5.6	2.9	3.6	1.3	versicolor	0	0	0
66	6.7	3.1	4.4	1.4	versicolor	0	0	0
67	5.6	3	4.5	1.5	versicolor	0	0	0
68	5.8	2.7	4.1	1	versicolor	0	0	0
69	6.2	2.2	4.5	1.5	versicolor	0	0	0
70	5.6	2.5	3.9	1.1	versicolor	0	0	0
71	5.9	3.2	4.8	1.8	versicolor	0	0	0
72	6.1	2.8	4	1.3	versicolor	0	0	0
73	6.3	2.5	4.9	1.5	versicolor	0	0	0
74	6.1	2.8	4.7	1.2	versicolor	0	0	0
75	6.4	2.9	4.3	1.3	versicolor	0	0	0
76	6.6	3	4.4	1.4	versicolor	0	0	0
77	6.8	2.8	4.8	1.4	versicolor	0	0	0
78	6.7	3	5	1.7	versicolor	0	0	0
79	6	2.9	4.5	1.5	versicolor	0	0	0
80	5.7	2.6	3.5	1	versicolor	0	0	0
81	5.5	2.4	3.8	1.1	versicolor	0	0	0
82	5.5	2.4	3.7	1	versicolor	0	0	0
83	5.8	2.7	3.9	1.2	versicolor	0	0	0
84	6	2.7	5.1	1.6	versicolor	0	0	0
85	5.4	3	4.5	1.5	versicolor	0	0	0
86	6	3.4	4.5	1.6	versicolor	0	0	0
87	6.7	3.1	4.7	1.5	versicolor	0	0	0
88	6.3	2.3	4.4	1.3	versicolor	0	0	0
89	5.6	3	4.1	1.3	versicolor	0	0	0
90	5.5	2.5	4	1.3	versicolor	0	0	0
91	5.5	2.6	4.4	1.2	versicolor	0	0	0
92	6.1	3	4.6	1.4	versicolor	0	0	0
93	5.8	2.6	4	1.2	versicolor	0	0	0
94	5	2.3	3.3	1	versicolor	0	0	0
95	5.6	2.7	4.2	1.3	versicolor	0	0	0
96	5.7	3	4.2	1.2	versicolor	0	0	0
97	5.7	2.9	4.2	1.3	versicolor	0	0	0
98	6.2	2.9	4.3	1.3	versicolor	0	0	0
99	5.1	2.5	3	1.1	versicolor	0	0	0
100	5.7	2.8	4.1	1.3	versicolor	0	0	0
101	6.3	3.3	6	2.5	virginica	0	0	0
102	5.8	2.7	5.1	1.9	virginica	0	0	0
103	7.1	3	5.9	2.1	virginica	0	0	0
104	6.3	2.9	5.6	1.8	virginica	0	0	0
105	6.5	3	5.8	2.2	virginica	0	0	0

106	7.6	3	6.6	2.1	virginica	0	0	0
107	4.9	2.5	4.5	1.7	virginica	0	0	0
108	7.3	2.9	6.3	1.8	virginica	0	0	0
109	6.7	2.5	5.8	1.8	virginica	0	0	0
110	7.2	3.6	6.1	2.5	virginica	0	0	0
111	6.5	3.2	5.1	2	virginica	0	0	0
112	6.4	2.7	5.3	1.9	virginica	0	0	0
113	6.8	3	5.5	2.1	virginica	0	0	0
114	5.7	2.5	5	2	virginica	0	0	0
115	5.8	2.8	5.1	2.4	virginica	0	0	0
116	6.4	3.2	5.3	2.3	virginica	0	0	0
117	6.5	3	5.5	1.8	virginica	0	0	0
118	7.7	3.8	6.7	2.2	virginica	0	0	0
119	7.7	2.6	6.9	2.3	virginica	0	0	0
120	6	2.2	5	1.5	virginica	0	0	0
121	6.9	3.2	5.7	2.3	virginica	0	0	0
122	5.6	2.8	4.9	2	virginica	0	0	0
123	7.7	2.8	6.7	2	virginica	0	0	0
124	6.3	2.7	4.9	1.8	virginica	0	0	0
125	6.7	3.3	5.7	2.1	virginica	0	0	0
126	7.2	3.2	6	1.8	virginica	0	0	0
127	6.2	2.8	4.8	1.8	virginica	0	0	0
128	6.1	3	4.9	1.8	virginica	0	0	0
129	6.4	2.8	5.6	2.1	virginica	0	0	0
130	7.2	3	5.8	1.6	virginica	0	0	0
131	7.4	2.8	6.1	1.9	virginica	0	0	0
132	7.9	3.8	6.4	2	virginica	0	0	0
133	6.4	2.8	5.6	2.2	virginica	0	0	0
134	6.3	2.8	5.1	1.5	virginica	0	0	0
135	6.1	2.6	5.6	1.4	virginica	0	0	0
136	7.7	3	6.1	2.3	virginica	0	0	0
137	6.3	3.4	5.6	2.4	virginica	0	0	0
138	6.4	3.1	5.5	1.8	virginica	0	0	0
139	6	3	4.8	1.8	virginica	0	0	0
140	6.9	3.1	5.4	2.1	virginica	0	0	0
141	6.7	3.1	5.6	2.4	virginica	0	0	0
142	6.9	3.1	5.1	2.3	virginica	0	0	0
143	5.8	2.7	5.1	1.9	virginica	0	0	0
144	6.8	3.2	5.9	2.3	virginica	0	0	0
145	6.7	3.3	5.7	2.5	virginica	0	0	0
146	6.7	3	5.2	2.3	virginica	0	0	0
147	6.3	2.5	5	1.9	virginica	0	0	0
148	6.5	3	5.2	2	virginica	0	0	0
149	6.2	3.4	5.4	2.3	virginica	0	0	0
150	5.9	3	5.1	1.8	virginica	0	0	0

Define a variable x2 which takes the value 2 when the species is versicolor or 0 otherwise. Performing binary logistic on x2 using the given independent variables we get,

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sepal_Length	-.245	.650	.143	1	.706	.782
	Sepal_Width	-2.797	.784	12.739	1	.000	.061
	Petal_Length	1.314	.684	3.691	1	.055	3.720
	Petal_Width	-2.778	1.173	5.609	1	.018	.062
	Constant	7.378	2.499	8.716	1	.003	1601.165

a. Variable(s) entered on step 1: Sepal_Length, Sepal_Width, Petal_Length, Petal_Width.

The p-value for the test of significance is less than the level of significance (0.05) for the variables- Sepal Width, Petal Width. Hence these variables are significant whereas the other two are non significant.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.524	8	.384

The p-value of the Hosmer Lemeshow test is 0.384, which is greater than the level of significance; hence the null hypothesis is accepted. The fitted model is good model for the data at hand.

Classification Table[a]

Observed		Predicted		
		m2		Percentage Correct
		.00	2.00	
Step 1	m2 .00	86	14	86.0
	2.00	25	25	50.0
	Overall Percentage			74.0

a. The cut value is .500

The percentage of correct classification is 74%, which suggests that the model is a somewhat good model in classifying the data.

- Sensitivity = 25/(25+25) = 0.50
- Specificity = 86/(86+14) = 0.86

Predicted probabilities and predicted group for x2 is given below (using Binary Logistic):

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	x2	Pre_prob2	Pre_group2
1	5.1	3.5	1.4	0.2	setosa	0	0.08491	02
2	4.9	3	1.4	0.2	setosa	0	0.28292	0
3	4.7	3.2	1.3	0.2	setosa	0	0.17198	0
4	4.6	3.1	1.5	0.2	setosa	0	0.26801	0

5	5	3.6	1.4	0.2	setosa	0	0.06708	0
6	5.4	3.9	1.7	0.4	setosa	0	0.0234	0
7	4.6	3.4	1.4	0.3	setosa	0	0.0951	0
8	5	3.4	1.5	0.2	setosa	0	0.12545	0
9	4.4	2.9	1.4	0.2	setosa	0	0.37105	0
10	4.9	3.1	1.5	0.1	setosa	0	0.30992	0
11	5.4	3.7	1.5	0.2	setosa	0	0.0532	0
12	4.8	3.4	1.6	0.2	setosa	0	0.14662	0
13	4.8	3	1.4	0.1	setosa	0	0.34804	0
14	4.3	3	1.1	0.1	setosa	0	0.28924	0
15	5.8	4	1.2	0.2	setosa	0	0.01463	0
16	5.7	4.4	1.5	0.4	setosa	0	0.00421	0
17	5.4	3.9	1.3	0.4	setosa	0	0.01397	0
18	5.1	3.5	1.4	0.3	setosa	0	0.06567	0
19	5.7	3.8	1.7	0.3	setosa	0	0.03742	0
20	5.1	3.8	1.5	0.3	setosa	0	0.03348	0
21	5.4	3.4	1.7	0.2	setosa	0	0.14464	0
22	5.1	3.7	1.5	0.4	setosa	0	0.03354	0
23	4.6	3.6	1	0.2	setosa	0	0.0448	0
24	5.1	3.3	1.7	0.5	setosa	0	0.09471	0
25	4.8	3.4	1.9	0.2	setosa	0	0.20306	0
26	5	3	1.6	0.2	setosa	0	0.33362	0
27	5	3.4	1.6	0.4	setosa	0	0.08579	0
28	5.2	3.5	1.5	0.2	setosa	0	0.09359	0
29	5.2	3.4	1.4	0.2	setosa	0	0.10695	0
30	4.7	3.2	1.6	0.2	setosa	0	0.23549	0
31	4.8	3.1	1.6	0.2	setosa	0	0.28446	0
32	5.4	3.4	1.5	0.4	setosa	0	0.06942	0
33	5.2	4.1	1.5	0.1	setosa	0	0.02483	0
34	5.5	4.2	1.4	0.2	setosa	0	0.01174	0
35	4.9	3.1	1.5	0.2	setosa	0	0.25382	0
36	5	3.2	1.2	0.2	setosa	0	0.14472	0
37	5.5	3.5	1.3	0.2	setosa	0	0.0687	0
38	4.9	3.6	1.4	0.1	setosa	0	0.08866	0
39	4.4	3	1.3	0.2	setosa	0	0.28116	0
40	5.1	3.4	1.5	0.2	setosa	0	0.12278	0
41	5	3.5	1.3	0.3	setosa	0	0.05941	0
42	4.5	2.3	1.3	0.3	setosa	0	0.67184	1
43	4.4	3.2	1.3	0.2	setosa	0	0.18272	0
44	5	3.5	1.6	0.6	setosa	0	0.03911	0
45	5.1	3.8	1.9	0.4	setosa	0	0.04248	0

46	4.8	3	1.4	0.3	setosa	0	0.23445	0
47	5.1	3.8	1.6	0.2	setosa	0	0.04957	0
48	4.6	3.2	1.4	0.2	setosa	0	0.19533	0
49	5.3	3.7	1.5	0.2	setosa	0	0.05445	0
50	5	3.3	1.4	0.2	setosa	0	0.14264	0
51	7	3.2	4.7	1.4	versicolor	1	0.26824	0
52	6.4	3.2	4.5	1.5	versicolor	1	0.1983	0
53	6.9	3.1	4.9	1.5	versicolor	1	0.32861	0
54	5.5	2.3	4	1.3	versicolor	1	0.7755	1
55	6.5	2.8	4.6	1.5	versicolor	1	0.45723	0
56	5.7	2.8	4.5	1.3	versicolor	1	0.61043	1
57	6.3	3.3	4.7	1.6	versicolor	1	0.1588	0
58	4.9	2.4	3.3	1	versicolor	1	0.7352	1
59	6.6	2.9	4.6	1.3	versicolor	1	0.51999	1
60	5.2	2.7	3.9	1.4	versicolor	1	0.44656	0
61	5	2	3.5	1	versicolor	1	0.91513	1
62	5.9	3	4.2	1.5	versicolor	1	0.24805	0
63	6	2.2	4	1	versicolor	1	0.90292	1
64	6.1	2.9	4.7	1.4	versicolor	1	0.51405	1
65	5.6	2.9	3.6	1.3	versicolor	1	0.27125	0
66	6.7	3.1	4.4	1.4	versicolor	1	0.2603	0
67	5.6	3	4.5	1.5	versicolor	1	0.34494	0
68	5.8	2.7	4.1	1	versicolor	1	0.73347	1
69	6.2	2.2	4.5	1.5	versicolor	1	0.8098	1
70	5.6	2.5	3.9	1.1	versicolor	1	0.74651	1
71	5.9	3.2	4.8	1.8	versicolor	1	0.15269	0
72	6.1	2.8	4	1.3	versicolor	1	0.42413	0
73	6.3	2.5	4.9	1.5	versicolor	1	0.75225	1
74	6.1	2.8	4.7	1.2	versicolor	1	0.70921	1
75	6.4	2.9	4.3	1.3	versicolor	1	0.43413	0
76	6.6	3	4.4	1.4	versicolor	1	0.32296	0
77	6.8	2.8	4.8	1.4	versicolor	1	0.57334	1
78	6.7	3	5	1.7	versicolor	1	0.30787	0
79	6	2.9	4.5	1.5	versicolor	1	0.38703	0
80	5.7	2.6	3.5	1	versicolor	1	0.62908	1
81	5.5	2.4	3.8	1.1	versicolor	1	0.77781	1
82	5.5	2.4	3.7	1	versicolor	1	0.80209	1
83	5.8	2.7	3.9	1.2	versicolor	1	0.54832	1
84	6	2.7	5.1	1.6	versicolor	1	0.64791	1
85	5.4	3	4.5	1.5	versicolor	1	0.35612	0
86	6	3.4	4.5	1.6	versicolor	1	0.10565	0

87	6.7	3.1	4.7	1.5	versicolor	1	0.2833	0
88	6.3	2.3	4.4	1.3	versicolor	1	0.82761	1
89	5.6	3	4.1	1.3	versicolor	1	0.3518	0
90	5.5	2.5	4	1.3	versicolor	1	0.66381	1
91	5.5	2.6	4.4	1.2	versicolor	1	0.76923	1
92	6.1	3	4.6	1.4	versicolor	1	0.41222	0
93	5.8	2.6	4	1.2	versicolor	1	0.64678	1
94	5	2.3	3.3	1	versicolor	1	0.78182	1
95	5.6	2.7	4.2	1.3	versicolor	1	0.58885	1
96	5.7	3	4.2	1.2	versicolor	1	0.44363	0
97	5.7	2.9	4.2	1.3	versicolor	1	0.44408	0
98	6.2	2.9	4.3	1.3	versicolor	1	0.44622	0
99	5.1	2.5	3	1.1	versicolor	1	0.50512	1
100	5.7	2.8	4.1	1.3	versicolor	1	0.48092	0
101	6.3	3.3	6	2.5	virginica	0	0.07872	0
102	5.8	2.7	5.1	1.9	virginica	0	0.45646	0
103	7.1	3	5.9	2.1	virginica	0	0.30211	0
104	6.3	2.9	5.6	1.8	virginica	0	0.51951	1
105	6.5	3	5.8	2.2	virginica	0	0.24988	0
106	7.6	3	6.6	2.1	virginica	0	0.4899	0
107	4.9	2.5	4.5	1.7	virginica	0	0.59219	1
108	7.3	2.9	6.3	1.8	virginica	0	0.67967	1
109	6.7	2.5	5.8	1.8	virginica	0	0.79597	1
110	7.2	3.6	6.1	2.5	virginica	0	0.03266	0
111	6.5	3.2	5.1	2	virginica	0	0.11687	0
112	6.4	2.7	5.3	1.9	virginica	0	0.48524	0
113	6.8	3	5.5	2.1	virginica	0	0.216	0
114	5.7	2.5	5	2	virginica	0	0.50002	1
115	5.8	2.8	5.1	2.4	virginica	0	0.13665	0
116	6.4	3.2	5.3	2.3	virginica	0	0.07118	0
117	6.5	3	5.5	1.8	virginica	0	0.40564	0
118	7.7	3.8	6.7	2.2	virginica	0	0.07954	0
119	7.7	2.6	6.9	2.3	virginica	0	0.70933	1
120	6	2.2	5	1.5	virginica	0	0.8961	1
121	6.9	3.2	5.7	2.3	virginica	0	0.10285	0
122	5.6	2.8	4.9	2	virginica	0	0.27974	0
123	7.7	2.8	6.7	2	virginica	0	0.71168	1
124	6.3	2.7	4.9	1.8	virginica	0	0.42992	0
125	6.7	3.3	5.7	2.1	virginica	0	0.13695	0
126	7.2	3.2	6	1.8	virginica	0	0.38787	0
127	6.2	2.8	4.8	1.8	virginica	0	0.3388	0

128	6.1	3	4.9	1.8	virginica	0	0.25501	0
129	6.4	2.8	5.6	2.1	virginica	0	0.37747	0
130	7.2	3	5.8	1.6	virginica	0	0.59773	1
131	7.4	2.8	6.1	1.9	virginica	0	0.61463	1
132	7.9	3.8	6.4	2	virginica	0	0.08817	0
133	6.4	2.8	5.6	2.2	virginica	0	0.31472	0
134	6.3	2.8	5.1	1.5	virginica	0	0.63051	1
135	6.1	2.6	5.6	1.4	virginica	0	0.88869	1
136	7.7	3	6.1	2.3	virginica	0	0.21799	0
137	6.3	3.4	5.6	2.4	virginica	0	0.04801	0
138	6.4	3.1	5.5	1.8	virginica	0	0.34589	0
139	6	3	4.8	1.8	virginica	0	0.23525	0
140	6.9	3.1	5.4	2.1	virginica	0	0.15127	0
141	6.7	3.1	5.6	2.4	virginica	0	0.09566	0
142	6.9	3.1	5.1	2.3	virginica	0	0.0645	0
143	5.8	2.7	5.1	1.9	virginica	0	0.45646	0
144	6.8	3.2	5.9	2.3	virginica	0	0.13254	0
145	6.7	3.3	5.7	2.5	virginica	0	0.04963	0
146	6.7	3	5.2	2.3	virginica	0	0.09847	0
147	6.3	2.5	5	1.9	virginica	0	0.53262	1
148	6.5	3	5.2	2	virginica	0	0.20887	0
149	6.2	3.4	5.4	2.3	virginica	0	0.04985	0
150	5.9	3	5.1	1.8	virginica	0	0.31859	0

Define a variable x3 which takes the value 3 when the species is Virginia or 0 otherwise.

Run binary logistic on x3 using the given independent variables we get

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sepal_Length	-2.465	2.394	1.060	1	.303	.085
	Sepal_Width	-6.681	4.480	2.224	1	.136	.001
	Petal_Length	9.429	4.737	3.962	1	.047	12448.870
	Petal_Width	18.286	9.743	3.523	1	.061	8.741E7
	Constant	-42.638	25.708	2.751	1	.097	.000

a. Variable(s) entered on step 1: Sepal_Length, Sepal_Width, Petal_Length, Petal_Width.

The p-values for the test of significances is less than the level of significance for the variable Petal Length, which implies that it is significant, while the rest of the variables are all insignificant.

Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	.259	8	1.000

The p-value for the Hosmer and Lemeshow test is 1, which is greater than the level of significance. Hence, the null hypothesis is accepted which implies that the fitted binary logistic model is a good fit for the data at hand.

Classification Table[a]

Observed			Predicted		
			m3		Percentage Correct
			.00	3.00	
Step 1	m3	.00	99	1	99.0
		3.00	1	49	98.0
	Overall Percentage				98.7

a. The cut value is .500

The percentage of correct classification is 98.7, which is a high value implying that the model fitted is good model to predict the species as Virginica or not.

- Sensitivity = 49/(49+1) = 0.98
- Specificity = 99/(99+1) = 0.99

Predicted probabilities and predicted groups of x3 is given below (using Binary Logistic):

Flower no.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	x3	Pre_prob3	Pre_group3
1	5.1	3.5	1.4	0.2	setosa	0	0	0
2	4.9	3	1.4	0.2	setosa	0	0	0
3	4.7	3.2	1.3	0.2	setosa	0	0	0
4	4.6	3.1	1.5	0.2	setosa	0	0	0
5	5	3.6	1.4	0.2	setosa	0	0	0
6	5.4	3.9	1.7	0.4	setosa	0	0	0
7	4.6	3.4	1.4	0.3	setosa	0	0	0
8	5	3.4	1.5	0.2	setosa	0	0	0
9	4.4	2.9	1.4	0.2	setosa	0	0	0
10	4.9	3.1	1.5	0.1	setosa	0	0	0
11	5.4	3.7	1.5	0.2	setosa	0	0	0
12	4.8	3.4	1.6	0.2	setosa	0	0	0
13	4.8	3	1.4	0.1	setosa	0	0	0
14	4.3	3	1.1	0.1	setosa	0	0	0
15	5.8	4	1.2	0.2	setosa	0	0	0
16	5.7	4.4	1.5	0.4	setosa	0	0	0
17	5.4	3.9	1.3	0.4	setosa	0	0	0
18	5.1	3.5	1.4	0.3	setosa	0	0	0
19	5.7	3.8	1.7	0.3	setosa	0	0	0
20	5.1	3.8	1.5	0.3	setosa	0	0	0
21	5.4	3.4	1.7	0.2	setosa	0	0	0
22	5.1	3.7	1.5	0.4	setosa	0	0	0
23	4.6	3.6	1	0.2	setosa	0	0	0
24	5.1	3.3	1.7	0.5	setosa	0	0	0

25	4.8	3.4	1.9	0.2	setosa	0	0	0
26	5	3	1.6	0.2	setosa	0	0	0
27	5	3.4	1.6	0.4	setosa	0	0	0
28	5.2	3.5	1.5	0.2	setosa	0	0	0
29	5.2	3.4	1.4	0.2	setosa	0	0	0
30	4.7	3.2	1.6	0.2	setosa	0	0	0
31	4.8	3.1	1.6	0.2	setosa	0	0	0
32	5.4	3.4	1.5	0.4	setosa	0	0	0
33	5.2	4.1	1.5	0.1	setosa	0	0	0
34	5.5	4.2	1.4	0.2	setosa	0	0	0
35	4.9	3.1	1.5	0.2	setosa	0	0	0
36	5	3.2	1.2	0.2	setosa	0	0	0
37	5.5	3.5	1.3	0.2	setosa	0	0	0
38	4.9	3.6	1.4	0.1	setosa	0	0	0
39	4.4	3	1.3	0.2	setosa	0	0	0
40	5.1	3.4	1.5	0.2	setosa	0	0	0
41	5	3.5	1.3	0.3	setosa	0	0	0
42	4.5	2.3	1.3	0.3	setosa	0	0	0
43	4.4	3.2	1.3	0.2	setosa	0	0	0
44	5	3.5	1.6	0.6	setosa	0	0	0
45	5.1	3.8	1.9	0.4	setosa	0	0	0
46	4.8	3	1.4	0.3	setosa	0	0	0
47	5.1	3.8	1.6	0.2	setosa	0	0	0
48	4.6	3.2	1.4	0.2	setosa	0	0	0
49	5.3	3.7	1.5	0.2	setosa	0	0	0
50	5	3.3	1.4	0.2	setosa	0	0	0
51	7	3.2	4.7	1.4	versicolor	0	0	0
52	6.4	3.2	4.5	1.5	versicolor	0	0.0001	0
53	6.9	3.1	4.9	1.5	versicolor	0	0.0012	0
54	5.5	2.3	4	1.3	versicolor	0	0	0
55	6.5	2.8	4.6	1.5	versicolor	0	0.0014	0
56	5.7	2.8	4.5	1.3	versicolor	0	0.0001	0
57	6.3	3.3	4.7	1.6	versicolor	0	0.0013	0
58	4.9	2.4	3.3	1	versicolor	0	0	0
59	6.6	2.9	4.6	1.3	versicolor	0	0	0
60	5.2	2.7	3.9	1.4	versicolor	0	0	0
61	5	2	3.5	1	versicolor	0	0	0
62	5.9	3	4.2	1.5	versicolor	0	0	0
63	6	2.2	4	1	versicolor	0	0	0
64	6.1	2.9	4.7	1.4	versicolor	0	0.0008	0
65	5.6	2.9	3.6	1.3	versicolor	0	0	0

66	6.7	3.1	4.4	1.4	versicolor	0	0	0
67	5.6	3	4.5	1.5	versicolor	0	0.0013	0
68	5.8	2.7	4.1	1	versicolor	0	0	0
69	6.2	2.2	4.5	1.5	versicolor	0	0.0596	0
70	5.6	2.5	3.9	1.1	versicolor	0	0	0
71	5.9	3.2	4.8	1.8	versicolor	0	0.4048	0
72	6.1	2.8	4	1.3	versicolor	0	0	0
73	6.3	2.5	4.9	1.5	versicolor	0	0.2248	0
74	6.1	2.8	4.7	1.2	versicolor	0	0	0
75	6.4	2.9	4.3	1.3	versicolor	0	0	0
76	6.6	3	4.4	1.4	versicolor	0	0	0
77	6.8	2.8	4.8	1.4	versicolor	0	0.0007	0
78	6.7	3	5	1.7	versicolor	0	0.2761	0
79	6	2.9	4.5	1.5	versicolor	0	0.001	0
80	5.7	2.6	3.5	1	versicolor	0	0	0
81	5.5	2.4	3.8	1.1	versicolor	0	0	0
82	5.5	2.4	3.7	1	versicolor	0	0	0
83	5.8	2.7	3.9	1.2	versicolor	0	0	0
84	6	2.7	5.1	1.6	versicolor	0	0.8676	1
85	5.4	3	4.5	1.5	versicolor	0	0.0022	0
86	6	3.4	4.5	1.6	versicolor	0	0.0002	0
87	6.7	3.1	4.7	1.5	versicolor	0	0.0003	0
88	6.3	2.3	4.4	1.3	versicolor	0	0.0003	0
89	5.6	3	4.1	1.3	versicolor	0	0	0
90	5.5	2.5	4	1.3	versicolor	0	0	0
91	5.5	2.6	4.4	1.2	versicolor	0	0	0
92	6.1	3	4.6	1.4	versicolor	0	0.0002	0
93	5.8	2.6	4	1.2	versicolor	0	0	0
94	5	2.3	3.3	1	versicolor	0	0	0
95	5.6	2.7	4.2	1.3	versicolor	0	0	0
96	5.7	3	4.2	1.2	versicolor	0	0	0
97	5.7	2.9	4.2	1.3	versicolor	0	0	0
98	6.2	2.9	4.3	1.3	versicolor	0	0	0
99	5.1	2.5	3	1.1	versicolor	0	0	0
100	5.7	2.8	4.1	1.3	versicolor	0	0	0
101	6.3	3.3	6	2.5	virginica	1	1	1
102	5.8	2.7	5.1	1.9	virginica	1	0.9996	1
103	7.1	3	5.9	2.1	virginica	1	1	1
104	6.3	2.9	5.6	1.8	virginica	1	0.9997	1
105	6.5	3	5.8	2.2	virginica	1	1	1
106	7.6	3	6.6	2.1	virginica	1	1	1

107	4.9	2.5	4.5	1.7	virginica	1	0.8908	1
108	7.3	2.9	6.3	1.8	virginica	1	1	1
109	6.7	2.5	5.8	1.8	virginica	1	1	1
110	7.2	3.6	6.1	2.5	virginica	1	1	1
111	6.5	3.2	5.1	2	virginica	1	0.9903	1
112	6.4	2.7	5.3	1.9	virginica	1	0.9997	1
113	6.8	3	5.5	2.1	virginica	1	1	1
114	5.7	2.5	5	2	virginica	1	1	1
115	5.8	2.8	5.1	2.4	virginica	1	1	1
116	6.4	3.2	5.3	2.3	virginica	1	1	1
117	6.5	3	5.5	1.8	virginica	1	0.9977	1
118	7.7	3.8	6.7	2.2	virginica	1	1	1
119	7.7	2.6	6.9	2.3	virginica	1	1	1
120	6	2.2	5	1.5	virginica	1	0.9205	1
121	6.9	3.2	5.7	2.3	virginica	1	1	1
122	5.6	2.8	4.9	2	virginica	1	0.9995	1
123	7.7	2.8	6.7	2	virginica	1	1	1
124	6.3	2.7	4.9	1.8	virginica	1	0.9484	1
125	6.7	3.3	5.7	2.1	virginica	1	1	1
126	7.2	3.2	6	1.8	virginica	1	0.9996	1
127	6.2	2.8	4.8	1.8	virginica	1	0.8245	1
128	6.1	3	4.9	1.8	virginica	1	0.8023	1
129	6.4	2.8	5.6	2.1	virginica	1	1	1
130	7.2	3	5.8	1.6	virginica	1	0.9712	1
131	7.4	2.8	6.1	1.9	virginica	1	1	1
132	7.9	3.8	6.4	2	virginica	1	0.9999	1
133	6.4	2.8	5.6	2.2	virginica	1	1	1
134	6.3	2.8	5.1	1.5	virginica	1	0.2049	0
135	6.1	2.6	5.6	1.4	virginica	1	0.9664	1
136	7.7	3	6.1	2.3	virginica	1	1	1
137	6.3	3.4	5.6	2.4	virginica	1	1	1
138	6.4	3.1	5.5	1.8	virginica	1	0.9965	1
139	6	3	4.8	1.8	virginica	1	0.6691	1
140	6.9	3.1	5.4	2.1	virginica	1	0.9999	1
141	6.7	3.1	5.6	2.4	virginica	1	1	1
142	6.9	3.1	5.1	2.3	virginica	1	0.9999	1
143	5.8	2.7	5.1	1.9	virginica	1	0.9996	1
144	6.8	3.2	5.9	2.3	virginica	1	1	1
145	6.7	3.3	5.7	2.5	virginica	1	1	1
146	6.7	3	5.2	2.3	virginica	1	1	1
147	6.3	2.5	5	1.9	virginica	1	0.9991	1

148	6.5	3	5.2	2	virginica	1	0.999	1
149	6.2	3.4	5.4	2.3	virginica	1	1	1
150	5.9	3	5.1	1.8	virginica	1	0.9777	1

Classification matrix is given as:

Species * pred_spe Crosstabulation

Count

		pred_spe			
		setosa	versicolor	virginica	Total
Species	setosa	50	0	0	50
	versicolor	0	48	2	50
	virginica	0	1	49	50
Total		50	49	51	150

The percentage of correct classification is given as: [(50+48+49)/150]*100% = 98%. The model fitted seems to be a good fit for the given data.

Case 9: Consider the dataset **Flower Species.xlsx** dataset which has data on Sepal Length, Sepal Width, Petal Length and Petal Width and Species Type for 150 different flowers and perform the following objectives.

1. Built a multinomial logistic regression model for the given problem.
2. Obtain the predicted class for each flower using the multinomial logistic regression.
3. Construct the classification Matrix and report the percentage of correct classification.

H_o: All parameter of the i^{th} class is zero.

H_1: Atleast one of the parameter of the i^{th} class is non zero.

Likelihood Ratio Tests

	Model Fitting Criteria	Likelihood Ratio Tests		
Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	21.680	9.781	2	.008
SepalLength	13.266	1.367	2	.505
SepalWidth	15.492	3.594	2	.166
PetalLength	25.902	14.003	2	.001
PetalWidth	23.772	11.873	2	.003

From the table we can see that we will reject Ho in the case of petal_length and petal_width. Hence atleast one of the parameter of the class petal.length and petal.width is non zero.

Parameter Estimates

Species[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
setosa	Intercept	33.164	185175.457	.000	1	1.000			
	Sepal_Length	11.864	54562.262	.000	1	1.000	142063.517	.000	.[b]
	Sepal_Width	13.276	25968.390	.000	1	1.000	583190.925	.000	.[b]
	Petal_Length	-26.896	25481.604	.000	1	.999	2.086E-12	.000	.[b]
	Petal_Width	-38.067	.000	.	1	.	2.935E-17	2.935E-17	2.935E-17
versicolor	Intercept	42.638	25.708	2.751	1	.097			
	Sepal_Length	2.465	2.394	1.060	1	.303	11.766	.108	1284.293
	Sepal_Width	6.681	4.480	2.224	1	.136	797.026	.123	5181847.251
	Petal_Length	-9.429	4.737	3.962	1	.047	8.033E-5	7.457E-9	.865
	Petal_Width	-18.286	9.743	3.523	1	.061	1.144E-8	5.828E-17	2.246

a. The reference category is: virginica.

b. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

The intercepts for the model is given in the column B of the above table. The independent variables are significant when the p-value corresponding to the variable is less than 0.05.

SPSS uses flower type virginica as a reference and then compute the parameter estimates for flower type setosa and flower type versicolor.

Predicted class for each flower using the multinomial logistic regression

Flower No.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Pred_species
1	5.1	3.5	1.4	0.2	setosa	setosa
2	4.9	3	1.4	0.2	setosa	setosa
3	4.7	3.2	1.3	0.2	setosa	setosa
4	4.6	3.1	1.5	0.2	setosa	setosa
5	5	3.6	1.4	0.2	setosa	setosa
6	5.4	3.9	1.7	0.4	setosa	setosa
7	4.6	3.4	1.4	0.3	setosa	setosa
8	5	3.4	1.5	0.2	setosa	setosa
9	4.4	2.9	1.4	0.2	setosa	setosa
10	4.9	3.1	1.5	0.1	setosa	setosa
11	5.4	3.7	1.5	0.2	setosa	setosa
12	4.8	3.4	1.6	0.2	setosa	setosa
13	4.8	3	1.4	0.1	setosa	setosa

14	4.3	3	1.1	0.1	setosa	setosa
15	5.8	4	1.2	0.2	setosa	setosa
16	5.7	4.4	1.5	0.4	setosa	setosa
17	5.4	3.9	1.3	0.4	setosa	setosa
18	5.1	3.5	1.4	0.3	setosa	setosa
19	5.7	3.8	1.7	0.3	setosa	setosa
20	5.1	3.8	1.5	0.3	setosa	setosa
21	5.4	3.4	1.7	0.2	setosa	setosa
22	5.1	3.7	1.5	0.4	setosa	setosa
23	4.6	3.6	1	0.2	setosa	setosa
24	5.1	3.3	1.7	0.5	setosa	setosa
25	4.8	3.4	1.9	0.2	setosa	setosa
26	5	3	1.6	0.2	setosa	setosa
27	5	3.4	1.6	0.4	setosa	setosa
28	5.2	3.5	1.5	0.2	setosa	setosa
29	5.2	3.4	1.4	0.2	setosa	setosa
30	4.7	3.2	1.6	0.2	setosa	setosa
31	4.8	3.1	1.6	0.2	setosa	setosa
32	5.4	3.4	1.5	0.4	setosa	setosa
33	5.2	4.1	1.5	0.1	setosa	setosa
34	5.5	4.2	1.4	0.2	setosa	setosa
35	4.9	3.1	1.5	0.2	setosa	setosa
36	5	3.2	1.2	0.2	setosa	setosa
37	5.5	3.5	1.3	0.2	setosa	setosa
38	4.9	3.6	1.4	0.1	setosa	setosa
39	4.4	3	1.3	0.2	setosa	setosa
40	5.1	3.4	1.5	0.2	setosa	setosa
41	5	3.5	1.3	0.3	setosa	setosa
42	4.5	2.3	1.3	0.3	setosa	setosa
43	4.4	3.2	1.3	0.2	setosa	setosa
44	5	3.5	1.6	0.6	setosa	setosa
45	5.1	3.8	1.9	0.4	setosa	setosa
46	4.8	3	1.4	0.3	setosa	setosa
47	5.1	3.8	1.6	0.2	setosa	setosa
48	4.6	3.2	1.4	0.2	setosa	setosa
49	5.3	3.7	1.5	0.2	setosa	setosa
50	5	3.3	1.4	0.2	setosa	setosa
51	7	3.2	4.7	1.4	versicolor	versicolor
52	6.4	3.2	4.5	1.5	versicolor	versicolor
53	6.9	3.1	4.9	1.5	versicolor	versicolor
54	5.5	2.3	4	1.3	versicolor	versicolor

55	6.5	2.8	4.6	1.5	versicolor	versicolor
56	5.7	2.8	4.5	1.3	versicolor	versicolor
57	6.3	3.3	4.7	1.6	versicolor	versicolor
58	4.9	2.4	3.3	1	versicolor	versicolor
59	6.6	2.9	4.6	1.3	versicolor	versicolor
60	5.2	2.7	3.9	1.4	versicolor	versicolor
61	5	2	3.5	1	versicolor	versicolor
62	5.9	3	4.2	1.5	versicolor	versicolor
63	6	2.2	4	1	versicolor	versicolor
64	6.1	2.9	4.7	1.4	versicolor	versicolor
65	5.6	2.9	3.6	1.3	versicolor	versicolor
66	6.7	3.1	4.4	1.4	versicolor	versicolor
67	5.6	3	4.5	1.5	versicolor	versicolor
68	5.8	2.7	4.1	1	versicolor	versicolor
69	6.2	2.2	4.5	1.5	versicolor	versicolor
70	5.6	2.5	3.9	1.1	versicolor	versicolor
71	5.9	3.2	4.8	1.8	versicolor	versicolor
72	6.1	2.8	4	1.3	versicolor	versicolor
73	6.3	2.5	4.9	1.5	versicolor	versicolor
74	6.1	2.8	4.7	1.2	versicolor	versicolor
75	6.4	2.9	4.3	1.3	versicolor	versicolor
76	6.6	3	4.4	1.4	versicolor	versicolor
77	6.8	2.8	4.8	1.4	versicolor	versicolor
78	6.7	3	5	1.7	versicolor	versicolor
79	6	2.9	4.5	1.5	versicolor	versicolor
80	5.7	2.6	3.5	1	versicolor	versicolor
81	5.5	2.4	3.8	1.1	versicolor	versicolor
82	5.5	2.4	3.7	1	versicolor	versicolor
83	5.8	2.7	3.9	1.2	versicolor	versicolor
84	6	2.7	5.1	1.6	versicolor	Virginica
85	5.4	3	4.5	1.5	versicolor	Versicolor
86	6	3.4	4.5	1.6	versicolor	Versicolor
87	6.7	3.1	4.7	1.5	versicolor	Versicolor
88	6.3	2.3	4.4	1.3	versicolor	Versicolor
89	5.6	3	4.1	1.3	versicolor	Versicolor
90	5.5	2.5	4	1.3	versicolor	Versicolor
91	5.5	2.6	4.4	1.2	versicolor	Versicolor
92	6.1	3	4.6	1.4	versicolor	Versicolor
93	5.8	2.6	4	1.2	versicolor	Versicolor
94	5	2.3	3.3	1	versicolor	Versicolor
95	5.6	2.7	4.2	1.3	versicolor	Versicolor

96	5.7	3	4.2	1.2	versicolor	Versicolor
97	5.7	2.9	4.2	1.3	versicolor	Versicolor
98	6.2	2.9	4.3	1.3	versicolor	Versicolor
99	5.1	2.5	3	1.1	versicolor	Versicolor
100	5.7	2.8	4.1	1.3	versicolor	Versicolor
101	6.3	3.3	6	2.5	virginica	Virginica
102	5.8	2.7	5.1	1.9	virginica	Virginica
103	7.1	3	5.9	2.1	virginica	Virginica
104	6.3	2.9	5.6	1.8	virginica	virginica
105	6.5	3	5.8	2.2	virginica	virginica
106	7.6	3	6.6	2.1	virginica	virginica
107	4.9	2.5	4.5	1.7	virginica	virginica
108	7.3	2.9	6.3	1.8	virginica	virginica
109	6.7	2.5	5.8	1.8	virginica	virginica
110	7.2	3.6	6.1	2.5	virginica	virginica
111	6.5	3.2	5.1	2	virginica	virginica
112	6.4	2.7	5.3	1.9	virginica	virginica
113	6.8	3	5.5	2.1	virginica	virginica
114	5.7	2.5	5	2	virginica	virginica
115	5.8	2.8	5.1	2.4	virginica	virginica
116	6.4	3.2	5.3	2.3	virginica	virginica
117	6.5	3	5.5	1.8	virginica	virginica
118	7.7	3.8	6.7	2.2	virginica	virginica
119	7.7	2.6	6.9	2.3	virginica	virginica
120	6	2.2	5	1.5	virginica	virginica
121	6.9	3.2	5.7	2.3	virginica	virginica
122	5.6	2.8	4.9	2	virginica	virginica
123	7.7	2.8	6.7	2	virginica	virginica
124	6.3	2.7	4.9	1.8	virginica	virginica
125	6.7	3.3	5.7	2.1	virginica	virginica
126	7.2	3.2	6	1.8	virginica	virginica
127	6.2	2.8	4.8	1.8	virginica	virginica
128	6.1	3	4.9	1.8	virginica	virginica
129	6.4	2.8	5.6	2.1	virginica	virginica
130	7.2	3	5.8	1.6	virginica	virginica
131	7.4	2.8	6.1	1.9	virginica	virginica
132	7.9	3.8	6.4	2	virginica	virginica
133	6.4	2.8	5.6	2.2	virginica	virginica
134	6.3	2.8	5.1	1.5	virginica	versicolor
135	6.1	2.6	5.6	1.4	virginica	virginica
136	7.7	3	6.1	2.3	virginica	virginica

137	6.3	3.4	5.6	2.4	virginica	virginica
138	6.4	3.1	5.5	1.8	virginica	virginica
139	6	3	4.8	1.8	virginica	virginica
140	6.9	3.1	5.4	2.1	virginica	virginica
141	6.7	3.1	5.6	2.4	virginica	virginica
142	6.9	3.1	5.1	2.3	virginica	virginica
143	5.8	2.7	5.1	1.9	virginica	virginica
144	6.8	3.2	5.9	2.3	virginica	virginica
145	6.7	3.3	5.7	2.5	virginica	virginica
146	6.7	3	5.2	2.3	virginica	virginica
147	6.3	2.5	5	1.9	virginica	virginica
148	6.5	3	5.2	2	virginica	virginica
149	6.2	3.4	5.4	2.3	virginica	virginica
150	5.9	3	5.1	1.8	virginica	virginica

The classification table is given as:

Species * Predicted Response Category Crosstabulation

Count

		Predicted Response Category			
		setosa	versicolor	virginica	Total
Species	setosa	50	0	0	50
	versicolor	0	49	1	50
	virginica	0	1	49	50
Total		50	50	50	150

The percentage of correct classification is given as: [(50+49+49)/150]*100% = 98.67%. The model seems to be a good fit to the data at hand.

Case 10: Comparative Study of the two approaches of Multiclass Classification for the Flower Species Prediction Problem – Perform the following objectives:

1. Compare the predicted class for each flower based on both the approaches

2. Compare the diagonal entries of the classification matrix based on both the approaches.

3. Compare the percentage of correct classification for both the approaches.

Compare the predicted class for each flower based on both the approaches: Comparing from the predicted class of different flowers we conclude that

- Based on multiclass classification: In the data of 150 observations we have 3 mis-predicted values of flower types.

- Based on multinomial regression: In the data of 150 observations we have 2 mis-predicted values of flower types.

Hence definitely multinomial regression and multiclass classification approach are almost similar but multiclass classification is easier to understand.

Compare the diagonal entries of the classification matrix based on both the approaches

On comparing the diagonal entries of classification table in both the cases we observe that:

- For SETOSA in both the methods the models fit the data equally well
- For VERSICOLOR in multinomial regression the model fits the data better and
- For VERGINICA in both the methods the models fits the data equally well.

Compare the percentage of correct classification for both the approaches

- In multiclass classification percentage of correct classification for flower types s: setosa, versicolor, virginica are 100%, 96% and 98% respectively.
- In multinomial case the percentage of correct classification for flower types: setosa, versicolor, virginica are 100%, 98% and 98% respectively.

We conclude that the method of multiclass classification and multinomial regression are almost the same.

Sign- off Note:

Next step could be analyzing the separability to be able to select an appropriate classifier. Having appropriate classifiers how do we select the one with most confidence?