Case Study – Classification Problem

Submitted by: Vishal Kumar

Due Date: 19th October, 2015

Date of Submission: 16th October, 2015

Supervisor's Remarks

Late Submission:

Plagiarism:

Completeness:

Quality of Content:

Results and Interpretations:

Additional Remarks:

Email Spam Filtering

Email spam, also known as **junk email** or **unsolicited bulk email** is a mail involving nearly identical messages sent to numerous recipients by email. The messages may contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware. Now we want to develop a technique to filter out these spam emails.

An email will have some information in itself about the fact that its spam or not. That information we have to find out 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 number of persons to whom it has been sent, the name of the persons to whom it has been sent, the carbon copy list, the blind carbon copy list (which we cannot 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.

Naïve Bayes Classifier

Naïve Bayes classifier is a general technique for classification, here we will use it in the context of spam filtering. It is clear that the words of the message will tell us about mail 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.

Naïve Bayes classifier works in 3 steps process:

1. Computing the probability that the message is spam, knowing that a given word appears in this message.
2. Computing the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them).
3. 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 email 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:

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

P(S|W) is the probability that a message is a spam, knowing that the word ‘replica’ is in it.

(S) is the overall probability that any given message is spam.

($W|S$) is the probability that the word ‘replica’ appears in spam messages.

(H) is the overall probability that any given message is not spam (is "ham").

($W|H$) 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:

(S) = 0.80 and (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.

(S) = 0.50 and (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.

($W|S$) used in this formula is approximated to the frequency of messages containing ‘replica’ in the messages identified as spam during the learning phase.

($W|H$) is approximated to the frequency of messages containing ‘replica’ in the messages identified as ham during the learning phase. For these approximations to make sense, the set of learned messages needs to be big and representative enough. It is also advisable that the learned set of messages conforms to the 50% hypothesis about repartition between spam and ham, i.e. that the datasets of spam and ham are of same size. How this process does actually takes place? Suppose we have 500 spam emails and 500 ham emails (approximately half) we need to find ($W|S$) then we can write it as follows:

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

where $P(W \cap S)$ is calculated as the relative frequency of the word W in the spam messages. (S) is taken a priori. Similarly we can calculate ($H|S$).

Of course, determining whether a message is spam or ham based only on the presence of the word ‘replica’ is error-prone, which is why Bayesian spam software tries to consider several words and combine their spamicities to determine a message's overall probability of being spam.

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

Most of the 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:

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

p is the probability that the suspect message is spam.

p_1 is the probability ($S|W_1$) that it is a spam knowing it contains a first word (for example 'replica').

p_2 is the probability ($S|W_2$) that it is a spam knowing it contains a second word (for example 'watches').

p_n is the probability ($S|W_n$) that it is a spam knowing it contains an n^{th} word (for example 'home').

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.

Given P(S) = 0.5 and P(H) = 0.5, we will obtain P_i and 1-P_i. They are calculated in the table given below.

Word	**P(Word | Spam)**	**P(Word | Ham)**	**p_i**	**1-p_i**
Congratulations	0.8	0.2	0.800	0.200
winning	0.7	0.4	0.636	0.364
$	0.9	0.2	0.818	0.182
100000000	0.7	0.1	0.875	0.125
lottery	0.6	0.2	0.750	0.250
claim	0.6	0.3	0.667	0.333
prize	0.6	0.4	0.600	0.400
send	0.5	0.5	0.500	0.500
you	0.7	0.3	0.700	0.300
contact	0.5	0.5	0.500	0.500
details	0.6	0.6	0.500	0.500
Everything	0.2	0.7	0.222	0.778
going	0.2	0.7	0.222	0.778
fine	0.7	0.5	0.583	0.417
I	0.5	0.5	0.500	0.500
coming	0.5	0.5	0.500	0.500
summer	0.6	0.6	0.500	0.500
holidays	0.8	0.4	0.667	0.333
Take	0.7	0.6	0.538	0.462
care	0.2	0.2	0.500	0.500
yourself	0.8	0.7	0.533	0.467

Conclusion:

The overall spamicity of Email 1 is: **0.999784**, hence we classify this Email to be a spam message.

The overall spamicity of Email 2 is: **0.233577**, hence we classify this Email to be a ham.

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. We want to know "how more likely is a person to have cancer if he/she smokes rather he/she does not". 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. The data looks like:

Lung Cancer (Y)	Smoking (X)
1	0
0	0
1	1
0	1

Conti…

Building a logistic regression model for cancer possibility

We will run a logistic regression in SPSS and obtain the following table:

Variables in the Equation

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

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

So we have obtained the fitted logistic regression model as:

$$\widehat{\pi_i} = \frac{e^{-0.182+0.770x_1}}{1 + e^{-0.182+0.770x_1}}$$

Test for the Significance of Independent Variable

We see that significance value of "Smoking" variable is > 0.05, hence it is not a significant predictor at this level.

Confusion (Classification) Table

Confusion matrix, also known as a contingency table. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another).

Observed		Predicted		
		Lung Cancer		
		No Lung Cancer	Lung Cancer Present	Percentage Correct
Lung Cancer	No Lung Cancer	6	5	54.5
	Lung Cancer Present	5	9	64.3
Overall Percentage				60.0

a. The cut value is .500

Percentage of correct classification: 60%

Sensitivity: $\frac{9}{9+5} = 0.6428\ or\ 64.28\%$ i.e. 64% of the people were correctly identified to have lung cancer.

Specificity: $\frac{6}{6+5} = 0.5454\ or\ 54.54\%$ i.e. 54% of the people were correctly identified to have NO lung cancer.

Prediction of Cancer using the Logistic Classifier

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

Odds Ratio

OR: $e^{\boldsymbol{\beta_1}} = e^{0.77} = 2.519$

Odds of having cancer in the smoking group is 151% higher, than odds of having cancer in the non-smoking group.

Skull Type Prediction

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.

Based on the data given we will fit a logistic regression.

Building a 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.

$$\hat{\pi}_i = \frac{e^{1.149-0.008x_1-0.047x_2-0.007x_3-0.006x_4+0.022x_5}}{1+\ e^{1.149-0.008x_1-0.047x_2-0.007x_3-0.006x_4+0.022x_5}}$$

From the regression table we see that the p-value of all the regression coefficients are > 0.05, hence all the predictor variables are not significant at this level.

Overall Regression Test using Hosmer and Lemeshow Test

Hosmer and Lemeshow Test

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

Contingency Table for Hosmer and Lemeshow Test

		Skull_type = .00		Skull_type = 1.00		
		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

Conclusion:

We see the p-value is > 0.05, hence we conclude that model is a good fit.

Confusion (Classification) Matrix

Observed		Predicted		
		Skull_type		Percentage Correct
		.00	1.00	
Skull_type	.00	9	2	81.8
	1.00	3	5	62.5
Overall Percentage				73.7

Percentage of correct classification: 73.7% of the Skulls were correctly identified.

Sensitivity: $\frac{5}{5+3} = 0.625$ or 62.5% of the Type 2 Skulls were correctly identified.

Specificity: $\frac{9}{9+2} = 0.818$ or 81.8% of the Type 1 Skulls were correctly identified.

Prediction of Skull type using Logistic Classifier

Skull Type	X_1	X_2	X_3	X_4	X_5	Predicted Probability	Predicted Group
1	13	29	82	24	60	0.58476	1
0	79	1	1	56	63	0.82867	1
0	5	77	47	45	26	0.07618	0
1	100	16	80	60	98	0.72021	1
0	55	65	3	20	2	0.08238	0
0	91	55	20	59	68	0.25299	0
1	47	31	31	52	19	0.32244	0
1	17	43	61	45	79	0.52374	1
1	30	54	11	83	60	0.3039	0
0	45	17	63	79	10	0.34643	0
0	1	40	97	51	94	0.6031	1
1	69	44	40	47	84	0.47311	0
1	95	19	33	2	53	0.61877	1
0	83	47	75	68	9	0.08545	0
0	78	57	86	19	88	0.30121	0
1	94	1	50	44	88	0.85355	1
0	7	54	15	23	67	0.4543	0
0	77	45	34	32	75	0.42803	0
0	30	37	81	92	3	0.14077	0

For a set of five physical measures given for a new skull:

Skull Type	X_1	X_2	X_3	X_4	X_5
Predict?	171	134	130	69	130

Substituting these values in our estimated logistic model, we get

$$\widehat{\pi_i} = \frac{e^{1.149-0.008(171)-0.047(134)-0.007(130)-0.006(69)+0.022(130)}}{1 + e^{1.149-0.008(171)-0.047(134)-0.007(130)-0.006(69)+0.022(130)}}$$

$$\widehat{\pi_i} = 0.006867$$

Which is < 0.5, therefore skull is of **Type I**.

Sentiment Analysis using Logistic Regression

What makes a US Presidential Candidate Win?

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 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.

The first aspect of the problem is to understand the data itself, past win/loss statistics and the corresponding speeches. Clearly the response variable will indicate the win/loss information. 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.

1. Some measure indicating the content of speech showing *Openness.*
2. Some measure indicating the content of speech showing *Conscientiousness.*
3. Some measure indicating the content of speech showing *Extraversion.*
4. Some measure indicating the content of speech showing *Agreeableness.*
5. Some measure indicating the content of speech showing *Neuroticism.*
6. 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.

Logistic Regression Model for Classifying Win/Loss

Using SPSS we fit our model as:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Optimism(X_1)	-3.567	2.090	2.912	1	.088	.028
	Pessimism(X_2)	-28.451	2.951	92.952	1	.000	.000
	PastUsed(X_3)	2.080	.763	7.439	1	.006	8.002
	FutureUsed(X_4)	4.138	.725	32.559	1	.000	62.668
	OwnPartyCount(X_5)	.008	.006	1.756	1	.185	1.008
	OppPartyCount(X_6)	.016	.013	1.466	1	.226	1.016
	NumericContent(X_7)	311.989	53.759	33.680	1	.000	3.128E+135
	Extra(X_8)	-.377	.082	20.967	1	.000	.686
	Emoti(X_9)	.212	.130	2.669	1	.102	1.237
	Agree(X_{10})	-.583	.188	9.649	1	.002	.558
	Consc(X_{11})	-.481	.118	16.770	1	.000	.618
	Openn(X_{12})	.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.

Our model will be:

$$\hat{\pi_i} = \frac{e^{0.936-3.567x_1-28.451x_2-2.08x_3-4.138x_4+0.008x_5+0.016x_6+311.989x_7-0.377x_8+0.212x_9-0.583x_{10}-0.481x_{11}+0.828x_{12}}}{1+\ e^{0.936-3.567x_1-28.451x_2-2.08x_3-4.138x_4+0.008x_5+0.016x_6+311.989x_7-0.377x_8+0.212x_9-0.583x_{10}-0.481x_{11}+0.828x_{12}}}$$

We see that, based on p-value, variables like “Pessimisms”, “PastUsed”, “FutureUsed”, “NumericContent”, “Extra”, “Agree”, “Consc” and “Openn” have significant effect on Win/Loss of a candidate.

Overall Regression Test using Hosmer and Lemeshow Test

Hosmer and Lemeshow Test

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

Contingency Table for Hosmer and Lemeshow Test

		Win/Loss = Loss		Win/Loss = Win		
		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

Conclusion: We see the p-value is < 0.05, hence we conclude that model is not a good fit.

Confusion (Classification) Matrix

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

The cut value is .500

Percentage of correct classification: Winning status of 73.9% candidates were correctly identified.

Sensitivity: 83.1% of the winning candidates were correctly identified.

Specificity: 59.5% of the candidates who lost the elections were correctly identified.

Prediction of Win/Loss Status Using Logistic Classifier

Since our data set is too large, we have shown few predictions on winning/losing election of the candidates, in the table given below:

S.no.	Observed(Win/Loss)	Predictive Probability	Predicted(Win/Loss)
1	1	0.72361	1
2	1	0.75652	1
3	1	0.81174	1
4	1	0.73705	1
5	1	0.79277	1
6	1	0.84662	1
7	1	0.74538	1
8	1	0.66512	1
…	…	…	…
1516	0	0.56289	1
1517	0	0.26218	0
1518	0	0.35979	0
1519	0	0.2437	0
1520	0	0.85238	1
1521	0	0.48595	0
1522	0	0.23828	0
1523	0	0.99161	1
1524	0	0.3019	0

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 and π_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 and π_2. The aim is to find some "optimal" discriminant function. One such optimal rule is given by fisher linear discriminant function.

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,

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

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,

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

Using Cauchy Schwartz Inequality,

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

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

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 we deal with how to classify an arbitrary observation into one of the two possible classes.

Classification:

Realize that,

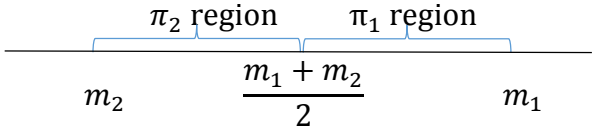
$$E[l'X \mid \pi_i] = E[(\mu_1 - \mu_2)'\Sigma^{-1}X \mid \pi_i] = (\mu_1 - \mu_2)'\Sigma^{-1}\mu_i \coloneqq m_i \; (say)$$

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 $y_0 = (\mu_1 - \mu_2)'\Sigma^{-1}x_0 = \text{Fisher LDF of } x_0$, 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,



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:

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

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

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

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

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:

if $(\widehat{\mu_1} - \widehat{\mu_2})'\hat{\Sigma}^{-1}x_0 > \frac{1}{2}(\widehat{\mu_1} - \widehat{\mu_2})'\hat{\Sigma}^{-1}(\widehat{\mu_1} + \widehat{\mu_2})$ then class is π_1 otherwise π_2.

The FLDF of x_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

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_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.

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.

H_0: Covariance matrices of the 2 groups do not differ significantly.
H_1: 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 neighbourhood of 1 or -1, then it is a good explanatory & if it lies in the

neighbourhood 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.

Test for Normality of All Five Physical Measures

We will test normality of all five physical measure by Kolmogorov-Smirnov:

H_0: Distribution of the physical measure is normal.

H_1: Distribution of the physical measure is not normal.

One-Sample Kolmogorov-Smirnov Test

		X1	X2	X3	X4	X5
N		19	19	19	19	19
Normal Parameters[a,b]	Mean	53.47	38.53	47.89	47.42	55.05
	Std. Deviation	34.407	20.759	30.269	23.691	32.987
Most Extreme Differences	Absolute	.174	.112	.131	.127	.191
	Positive	.121	.090	.098	.102	.127
	Negative	-.174	-.112	-.131	-.127	-.191
Test Statistic		.174	.112	.131	.127	.191
Asymp. Sig. (2-tailed)		.133[c]	.200[c,d]	.200[c,d]	.200[c,d]	.066[c]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

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

Conclusion: Last row of the table given above represents p-value of Kolmogorov-Smirnov test for all 5 physical measures. We find that for all measures p-value is > 0.05, hence we conclude that these measures follow normal distribution, at this level.

Test For Equality of the Covariance Matrices of Different Groups

We will test this by Box's M Test.

H_0: Covariance matrices of the different groups do not differ significantly.

H_1: Covariance matrices of the different groups differ significantly.

Conclusion: Since p-value is > 0.05, we conclude that, covariates matrices of the different groups do not 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.

Statistical Significance of Group Means

This is tested by Wilk's Lambda:

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

H_1: All the group means are 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

Conclusion: We see that all the variables have Wilk's Lambda value closer to 1, we conclude these variables do not contribute significantly to the discriminant function.

Overall Significance of the Discriminant Function

H_0: The overall discriminant function is not significant.

H_1: The overall discriminant function is significant.

Wilks' Lambda

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

Conclusion: Since p-value is > 0.05, we fail to reject H_0, hence we conclude that the overall discriminant function is not significant.

Standardized and Unstandardized Canonical Discriminant Function Coefficients

Standardized Canonical Discriminant Function Coefficients

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

Unstandardized Canonical Discriminant Function Coefficients

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

Conclusion: From unstandardized Canonical Discriminant Function Coefficient table we see, X_2 is a good explanatory variable as its value is close to 1, while others are moderate explanatory variables, as these values are close to -0.5 to 0.5.

Values of Discriminant Function at the Group Centroids and Classification Function Coefficients

Functions at Group Centroids

Skull Type	Function 1
0	.433
1	-.596

Unstandardized canonical discriminant functions evaluated at group means

Conclusion: From the table given above, $m_1 = 0.433$ and $m_2 = -0.596$.

Classification Function Coefficients

	Skull Type 0	Skull Type 1
X1	.100	.093
X2	.233	.189
X3	.055	.050
X4	.139	.132
X5	.062	.081
(Constant)	-14.614	-13.093

Fisher's linear discriminant functions

Discriminant Functions for the Given Skulls

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

FDLF $= l'X = 0.007X_1 + 0.044X_2 + 0.005X_3 + 0.007X_4 - 0.019X_5$

Predicted skull type for the given skulls:

Skull Type	X_1	X_2	X_3	X_4	X_5	Predicted Group	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

Classification Tables for Fitting Strength

Classification Results[a,c]

		Skull Type	Predicted Group Membership 0	Predicted Group Membership 1	Total
Original	Count	0	7	4	11
		1	2	6	8
	%	0	63.6	36.4	100.0
		1	25.0	75.0	100.0
Cross-validated[b]	Count	0	5	6	11
		1	5	3	8
	%	0	45.5	54.5	100.0
		1	62.5	37.5	100.0

a. 68.4% of original grouped cases correctly classified.

b. 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.

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

	Fitting Strength	**Predictive Strength**
Sensitivity	0.75	0.375
Specificity	0.636	0.455

Predict the Skull Type Using the Logistic Classifier

Skull Type	X_1	X_2	X_3	X_4	X_5
Predict?	171	134	130	69	130

FDLF = l'X = $0.007X_1 + 0.044X_2 + 0.005X_3 + 0.007X_4 - 0.019X_5$

If FDLF is > $\frac{m_1+m_2}{2}$ skull is Type 1 and if < $\frac{m_1+m_2}{2}$ skull is Type 2

FDLF = 0.007(171) + 0.044(134) + 0.005(130) + 0.007(69) – 0.019(130) = 5.756

$$\frac{m_1 + m_2}{2} = -0.0185$$

Conclusion: FDLF > -0.0185, hence Skull Type is 0(Group 1).

Multiclass Classification

A classification problem is said to be multiclass classification problem if the response has more than two possible classes. We will deal with 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:

1. By means of decomposing the multiclass classification problem into several binary classification problems.
2. 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:

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

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:

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

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 function 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. We will do the computational part through SPSS.

Multiclass Classification

We have considered the Flower Species dataset which has data on Sepal Length, Sepal Width, Petal Length, Petal Width and Species Type for 150 different flowers. We will decompose the multiclass (3-class) classification problem into three Binary Classification Problems and perform the analysis for each problem.

Case 1: "setosa" as 1 ,"versicolor" as 0 and "virginica" as 0.

Case 2: "setosa" as 0 ,"versicolor" as 1 and "virginica" as 0.

Case 3: "setosa" as 0 ,"versicolor" as 0 and "virginica" as 1.

Test For the Significance Individual Independent Variables

$\mathbf{H_0}$: $\beta_i = 0$ Vs $\mathbf{H_1}$: $\beta_i \neq 0$ for at least one i.

Case 1:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sepal.Length	8.666	14840.420	.000	1	1.000	5800.142
	Sepal.Width	6.637	6922.781	.000	1	.999	762.489
	Petal.Length	-15.119	12366.721	.000	1	.999	.000
	Petal.Width	-16.272	17915.189	.000	1	.999	.000
	Constant	-13.616	51859.147	.000	1	1.000	.000

a. Variable(s) entered on step 1: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width.

Conclusion: We fail to reject H_0 at 5% level of significance, hence we conclude β_i are not significant.

Case 2:

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.

Conclusion: Sepal width and Petal width are significant predictor at 5% level of significance.

Case 3:

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	87411453.756
	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.

Conclusion: Petal width is a significant predictor at 5% level of significance.

Test for the Overall Regression Significance

Case 1:

Hosmer and Lemeshow Test

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

Conclusion: Overall regression is not significant at 5% level of significance.

Case 2:

Hosmer and Lemeshow Test

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

Conclusion: Overall regression is not significant at 5% level of significance.

Case 3:

Hosmer and Lemeshow Test

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

Conclusion: Overall regression is not significant at 5% level of significance.

Confusion (Classification) Table

Case 1:

Classification Table[a]

Observed			Predicted		
			Setosa1		Percentage Correct
			0	1	
Step 1	Setosa1	0	100	0	100.0
		1	0	50	100.0
	Overall Percentage				100.0

a. The cut value is .500

Conclusion: Percentage of correct classification is 100%. Sensitivity is 100% and Specificity is 100%.

Case 2:

Classification Table[a]

Observed			Predicted		
			Versicolor1		Percentage Correct
			0	1	
Step 1	Versicolor1	0	86	14	86.0
		1	25	25	50.0
	Overall Percentage				74.0

a. The cut value is .500

Conclusion: Percentage of correct classification is 74%. Sensitivity is 50% and Specificity is 86%.

Case 3:

Classification Table[a]

Observed			Predicted		
			Virginica1		Percentage Correct
			0	1	
Step 1	Virginica1	0	99	1	99.0
		1	1	49	98.0
	Overall Percentage				98.7

a. The cut value is .500

Conclusion: Percentage of correct classification is 98.7%. Sensitivity is 98% and Specificity is 99%.

Predicted Probabilities and Predicted Groups

Flower No.	Species	Setosa Predicted Probability	Setosa Predicted Group	Versicolor Predicted Probability	Versicolor Predicted Group	Virginica Predicted Probability	Virginica Predicted Group
1	setosa	1.0000	1.0000	0.0849	0.0000	0.0000	0.0000
2	setosa	1.0000	1.0000	0.2829	0.0000	0.0000	0.0000
3	setosa	1.0000	1.0000	0.1720	0.0000	0.0000	0.0000
4	setosa	1.0000	1.0000	0.2680	0.0000	0.0000	0.0000
5	setosa	1.0000	1.0000	0.0671	0.0000	0.0000	0.0000
6	setosa	1.0000	1.0000	0.0234	0.0000	0.0000	0.0000
7	setosa	1.0000	1.0000	0.0951	0.0000	0.0000	0.0000
8	setosa	1.0000	1.0000	0.1255	0.0000	0.0000	0.0000
9	setosa	1.0000	1.0000	0.3711	0.0000	0.0000	0.0000
10	setosa	1.0000	1.0000	0.3099	0.0000	0.0000	0.0000
11	setosa	1.0000	1.0000	0.0532	0.0000	0.0000	0.0000
12	setosa	1.0000	1.0000	0.1466	0.0000	0.0000	0.0000
13	setosa	1.0000	1.0000	0.3480	0.0000	0.0000	0.0000
14	setosa	1.0000	1.0000	0.2892	0.0000	0.0000	0.0000
15	setosa	1.0000	1.0000	0.0146	0.0000	0.0000	0.0000
16	setosa	1.0000	1.0000	0.0042	0.0000	0.0000	0.0000
17	setosa	1.0000	1.0000	0.0140	0.0000	0.0000	0.0000
18	setosa	1.0000	1.0000	0.0657	0.0000	0.0000	0.0000
19	setosa	1.0000	1.0000	0.0374	0.0000	0.0000	0.0000
20	setosa	1.0000	1.0000	0.0335	0.0000	0.0000	0.0000
21	setosa	1.0000	1.0000	0.1446	0.0000	0.0000	0.0000
22	setosa	1.0000	1.0000	0.0335	0.0000	0.0000	0.0000
23	setosa	1.0000	1.0000	0.0448	0.0000	0.0000	0.0000
24	setosa	1.0000	1.0000	0.0947	0.0000	0.0000	0.0000
25	setosa	1.0000	1.0000	0.2031	0.0000	0.0000	0.0000
26	setosa	1.0000	1.0000	0.3336	0.0000	0.0000	0.0000
27	setosa	1.0000	1.0000	0.0858	0.0000	0.0000	0.0000
28	setosa	1.0000	1.0000	0.0936	0.0000	0.0000	0.0000
29	setosa	1.0000	1.0000	0.1070	0.0000	0.0000	0.0000
30	setosa	1.0000	1.0000	0.2355	0.0000	0.0000	0.0000
31	setosa	1.0000	1.0000	0.2845	0.0000	0.0000	0.0000
32	setosa	1.0000	1.0000	0.0694	0.0000	0.0000	0.0000
33	setosa	1.0000	1.0000	0.0248	0.0000	0.0000	0.0000
34	setosa	1.0000	1.0000	0.0117	0.0000	0.0000	0.0000
35	setosa	1.0000	1.0000	0.2538	0.0000	0.0000	0.0000
36	setosa	1.0000	1.0000	0.1447	0.0000	0.0000	0.0000
37	setosa	1.0000	1.0000	0.0687	0.0000	0.0000	0.0000
38	setosa	1.0000	1.0000	0.0887	0.0000	0.0000	0.0000
39	setosa	1.0000	1.0000	0.2812	0.0000	0.0000	0.0000
40	setosa	1.0000	1.0000	0.1228	0.0000	0.0000	0.0000
41	setosa	1.0000	1.0000	0.0594	0.0000	0.0000	0.0000

42	setosa	1.0000	1.0000	0.6718	1.0000	0.0000	0.0000
43	setosa	1.0000	1.0000	0.1827	0.0000	0.0000	0.0000
44	setosa	1.0000	1.0000	0.0391	0.0000	0.0000	0.0000
45	setosa	1.0000	1.0000	0.0425	0.0000	0.0000	0.0000
46	setosa	1.0000	1.0000	0.2345	0.0000	0.0000	0.0000
47	setosa	1.0000	1.0000	0.0496	0.0000	0.0000	0.0000
48	setosa	1.0000	1.0000	0.1953	0.0000	0.0000	0.0000
49	setosa	1.0000	1.0000	0.0545	0.0000	0.0000	0.0000
50	setosa	1.0000	1.0000	0.1426	0.0000	0.0000	0.0000
51	versicolor	0.0000	0.0000	0.2682	0.0000	0.0000	0.0000
52	versicolor	0.0000	0.0000	0.1983	0.0000	0.0001	0.0000
53	versicolor	0.0000	0.0000	0.3286	0.0000	0.0012	0.0000
54	versicolor	0.0000	0.0000	0.7755	1.0000	0.0000	0.0000
55	versicolor	0.0000	0.0000	0.4572	0.0000	0.0014	0.0000
56	versicolor	0.0000	0.0000	0.6104	1.0000	0.0001	0.0000
57	versicolor	0.0000	0.0000	0.1588	0.0000	0.0013	0.0000
58	versicolor	0.0000	0.0000	0.7352	1.0000	0.0000	0.0000
59	versicolor	0.0000	0.0000	0.5200	1.0000	0.0000	0.0000
60	versicolor	0.0000	0.0000	0.4466	0.0000	0.0000	0.0000
61	versicolor	0.0000	0.0000	0.9151	1.0000	0.0000	0.0000
62	versicolor	0.0000	0.0000	0.2481	0.0000	0.0000	0.0000
63	versicolor	0.0000	0.0000	0.9029	1.0000	0.0000	0.0000
64	versicolor	0.0000	0.0000	0.5141	1.0000	0.0008	0.0000
65	versicolor	0.0000	0.0000	0.2713	0.0000	0.0000	0.0000
66	versicolor	0.0000	0.0000	0.2603	0.0000	0.0000	0.0000
67	versicolor	0.0000	0.0000	0.3449	0.0000	0.0013	0.0000
68	versicolor	0.0000	0.0000	0.7335	1.0000	0.0000	0.0000
69	versicolor	0.0000	0.0000	0.8098	1.0000	0.0596	0.0000
70	versicolor	0.0000	0.0000	0.7465	1.0000	0.0000	0.0000
71	versicolor	0.0000	0.0000	0.1527	0.0000	0.4048	0.0000
72	versicolor	0.0000	0.0000	0.4241	0.0000	0.0000	0.0000
73	versicolor	0.0000	0.0000	0.7523	1.0000	0.2248	0.0000
74	versicolor	0.0000	0.0000	0.7092	1.0000	0.0000	0.0000
75	versicolor	0.0000	0.0000	0.4341	0.0000	0.0000	0.0000
76	versicolor	0.0000	0.0000	0.3230	0.0000	0.0000	0.0000
77	versicolor	0.0000	0.0000	0.5733	1.0000	0.0007	0.0000
78	versicolor	0.0000	0.0000	0.3079	0.0000	0.2761	0.0000
79	versicolor	0.0000	0.0000	0.3870	0.0000	0.0010	0.0000
80	versicolor	0.0000	0.0000	0.6291	1.0000	0.0000	0.0000
81	versicolor	0.0000	0.0000	0.7778	1.0000	0.0000	0.0000
82	versicolor	0.0000	0.0000	0.8021	1.0000	0.0000	0.0000
83	versicolor	0.0000	0.0000	0.5483	1.0000	0.0000	0.0000
84	versicolor	0.0000	0.0000	0.6479	1.0000	0.8676	1.0000
85	versicolor	0.0000	0.0000	0.3561	0.0000	0.0022	0.0000
86	versicolor	0.0000	0.0000	0.1057	0.0000	0.0002	0.0000
87	versicolor	0.0000	0.0000	0.2833	0.0000	0.0003	0.0000

88	versicolor	0.0000	0.0000	0.8276	1.0000	0.0003	0.0000
89	versicolor	0.0000	0.0000	0.3518	0.0000	0.0000	0.0000
90	versicolor	0.0000	0.0000	0.6638	1.0000	0.0000	0.0000
91	versicolor	0.0000	0.0000	0.7692	1.0000	0.0000	0.0000
92	versicolor	0.0000	0.0000	0.4122	0.0000	0.0002	0.0000
93	versicolor	0.0000	0.0000	0.6468	1.0000	0.0000	0.0000
94	versicolor	0.0000	0.0000	0.7818	1.0000	0.0000	0.0000
95	versicolor	0.0000	0.0000	0.5889	1.0000	0.0000	0.0000
96	versicolor	0.0000	0.0000	0.4436	0.0000	0.0000	0.0000
97	versicolor	0.0000	0.0000	0.4441	0.0000	0.0000	0.0000
98	versicolor	0.0000	0.0000	0.4462	0.0000	0.0000	0.0000
99	versicolor	0.0000	0.0000	0.5051	1.0000	0.0000	0.0000
100	versicolor	0.0000	0.0000	0.4809	0.0000	0.0000	0.0000
101	virginica	0.0000	0.0000	0.0787	0.0000	1.0000	1.0000
102	virginica	0.0000	0.0000	0.4565	0.0000	0.9996	1.0000
103	virginica	0.0000	0.0000	0.3021	0.0000	1.0000	1.0000
104	virginica	0.0000	0.0000	0.5195	1.0000	0.9997	1.0000
105	virginica	0.0000	0.0000	0.2499	0.0000	1.0000	1.0000
106	virginica	0.0000	0.0000	0.4899	0.0000	1.0000	1.0000
107	virginica	0.0000	0.0000	0.5922	1.0000	0.8908	1.0000
108	virginica	0.0000	0.0000	0.6797	1.0000	1.0000	1.0000
109	virginica	0.0000	0.0000	0.7960	1.0000	1.0000	1.0000
110	virginica	0.0000	0.0000	0.0327	0.0000	1.0000	1.0000
111	virginica	0.0000	0.0000	0.1169	0.0000	0.9903	1.0000
112	virginica	0.0000	0.0000	0.4852	0.0000	0.9997	1.0000
113	virginica	0.0000	0.0000	0.2160	0.0000	1.0000	1.0000
114	virginica	0.0000	0.0000	0.5000	1.0000	1.0000	1.0000
115	virginica	0.0000	0.0000	0.1367	0.0000	1.0000	1.0000
116	virginica	0.0000	0.0000	0.0712	0.0000	1.0000	1.0000
117	virginica	0.0000	0.0000	0.4056	0.0000	0.9977	1.0000
118	virginica	0.0000	0.0000	0.0795	0.0000	1.0000	1.0000
119	virginica	0.0000	0.0000	0.7093	1.0000	1.0000	1.0000
120	virginica	0.0000	0.0000	0.8961	1.0000	0.9205	1.0000
121	virginica	0.0000	0.0000	0.1029	0.0000	1.0000	1.0000
122	virginica	0.0000	0.0000	0.2797	0.0000	0.9995	1.0000
123	virginica	0.0000	0.0000	0.7117	1.0000	1.0000	1.0000
124	virginica	0.0000	0.0000	0.4299	0.0000	0.9484	1.0000
125	virginica	0.0000	0.0000	0.1370	0.0000	1.0000	1.0000
126	virginica	0.0000	0.0000	0.3879	0.0000	0.9996	1.0000
127	virginica	0.0000	0.0000	0.3388	0.0000	0.8245	1.0000
128	virginica	0.0000	0.0000	0.2550	0.0000	0.8023	1.0000
129	virginica	0.0000	0.0000	0.3775	0.0000	1.0000	1.0000
130	virginica	0.0000	0.0000	0.5977	1.0000	0.9712	1.0000
131	virginica	0.0000	0.0000	0.6146	1.0000	1.0000	1.0000
132	virginica	0.0000	0.0000	0.0882	0.0000	0.9999	1.0000
133	virginica	0.0000	0.0000	0.3147	0.0000	1.0000	1.0000

134	virginica	0.0000	0.0000	0.6305	1.0000	0.2049	0.0000
135	virginica	0.0000	0.0000	0.8887	1.0000	0.9664	1.0000
136	virginica	0.0000	0.0000	0.2180	0.0000	1.0000	1.0000
137	virginica	0.0000	0.0000	0.0480	0.0000	1.0000	1.0000
138	virginica	0.0000	0.0000	0.3459	0.0000	0.9965	1.0000
139	virginica	0.0000	0.0000	0.2353	0.0000	0.6691	1.0000
140	virginica	0.0000	0.0000	0.1513	0.0000	0.9999	1.0000
141	virginica	0.0000	0.0000	0.0957	0.0000	1.0000	1.0000
142	virginica	0.0000	0.0000	0.0645	0.0000	0.9999	1.0000
143	virginica	0.0000	0.0000	0.4565	0.0000	0.9996	1.0000
144	virginica	0.0000	0.0000	0.1325	0.0000	1.0000	1.0000
145	virginica	0.0000	0.0000	0.0496	0.0000	1.0000	1.0000
146	virginica	0.0000	0.0000	0.0985	0.0000	1.0000	1.0000
147	virginica	0.0000	0.0000	0.5326	1.0000	0.9991	1.0000
148	virginica	0.0000	0.0000	0.2089	0.0000	0.9990	1.0000
149	virginica	0.0000	0.0000	0.0499	0.0000	1.0000	1.0000
150	virginica	0.0000	0.0000	0.3186	0.0000	0.9777	1.0000

Multiclass Predicted Flower Species for the Original Problem Using Three Sub Problems

Flower No.	Observed Species	Predicted Species	Flower No.	Observed Species	Predicted Species	Flower No.	Observed Species	Predicted Species
1	1	1	51	2	2	101	3	3
2	1	1	52	2	2	102	3	3
3	1	1	53	2	2	103	3	3
4	1	1	54	2	2	104	3	3
5	1	1	55	2	2	105	3	3
6	1	1	56	2	2	106	3	3
7	1	1	57	2	2	107	3	3
8	1	1	58	2	2	108	3	3
9	1	1	59	2	2	109	3	3
10	1	1	60	2	2	110	3	3
11	1	1	61	2	2	111	3	3
12	1	1	62	2	2	112	3	3
13	1	1	63	2	2	113	3	3
14	1	1	64	2	2	114	3	3
15	1	1	65	2	2	115	3	3
16	1	1	66	2	2	116	3	3
17	1	1	67	2	2	117	3	3
18	1	1	68	2	2	118	3	3
19	1	1	69	2	2	119	3	3
20	1	1	70	2	2	120	3	3
21	1	1	71	2	3	121	3	3
22	1	1	72	2	2	122	3	3
23	1	1	73	2	2	123	3	3

24	1	1	74	2	2	124	3	3
25	1	1	75	2	2	125	3	3
26	1	1	76	2	2	126	3	3
27	1	1	77	2	2	127	3	3
28	1	1	78	2	2	128	3	3
29	1	1	79	2	2	129	3	3
30	1	1	80	2	2	130	3	3
31	1	1	81	2	2	131	3	3
32	1	1	82	2	2	132	3	3
33	1	1	83	2	2	133	3	3
34	1	1	84	2	3	134	3	2
35	1	1	85	2	2	135	3	3
36	1	1	86	2	2	136	3	3
37	1	1	87	2	2	137	3	3
38	1	1	88	2	2	138	3	3
39	1	1	89	2	2	139	3	3
40	1	1	90	2	2	140	3	3
41	1	1	91	2	2	141	3	3
42	1	1	92	2	2	142	3	3
43	1	1	93	2	2	143	3	3
44	1	1	94	2	2	144	3	3
45	1	1	95	2	2	145	3	3
46	1	1	96	2	2	146	3	3
47	1	1	97	2	2	147	3	3
48	1	1	98	2	2	148	3	3
49	1	1	99	2	2	149	3	3
50	1	1	100	2	2	150	3	3

Classification Matrix for Correct Classification

Predicted Species * Original Species Crosstabulation

Count

		Original Species			
		Setosa	Versicolor	Virginica	Total
Predicted Species	Setosa	50	0	0	50
	Versicolor	0	48	1	49
	Virginica	0	2	49	51
Total		50	50	50	150

Conclusion: Percentage of correct classification for species “Setosa” is 100%.

Percentage of correct classification for species “Versicolor” is 96%.

Percentage of correct classification for species “Virginica” is 98%.

Multinomial Logistic Regression

We have considered the Flower Species dataset which has data on Sepal Length, Sepal Width, Petal Length, Petal Width and Species Type for 150 different flowers. We will build a multinomial logistic regression and will further perform analysis on that.

Logistic Regression Model

Parameter Estimates

Species[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Versicolor	Intercept	6.756	5154.669	.000	1	.999			
	Sepal.Length	-6.615	2.394	7.633	1	.006	.001	1.228E-5	.146
	Sepal.Width	-5.450	1696.272	.000	1	.997	.004	.000	.[b]
	Petal.Length	12.856	770.161	.000	1	.987	382897.315	.000	.[b]
	Petal.Width	14.812	9.743	2.311	1	.128	2707999.265	.014	53156E+14
Virginica	Intercept	-35.882	5154.728	.000	1	.994			
	Sepal.Length	-9.080	.000	.	1	.	.000	.000	.000
	Sepal.Width	-12.130	1696.278	.000	1	.994	5.393E-6	.000	.[b]
	Petal.Length	22.285	770.183	.001	1	.977	4766638985	.000	.[b]
	Petal.Width	33.098	.000	.	1	.	2367E+14	2367E+14	2367E+14

a. The reference category is: Setosa.

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

Setosa was taken to be the reference group and model was fitted.

To test the significance of the model parameters we will do a **Likelihood Ratio Test**.

H_0: Parameter effect is zero.

H_1: Parameter effect is not zero.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	21.680[a]	9.781	2	.008
Sepal.Length	13.266[a]	1.367	2	.505
Sepal.Width	15.492[a]	3.594	2	.166
Petal.Length	25.902[a]	14.003	2	.001
Petal.Width	23.772[a]	11.873	2	.003

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Conclusion:

P-value is < 0.05 for Petal length and Sepal length, hence we reject H_0 at 5% level of significance and conclude that these 2 are significant predictor.

Predicted Class for Each Flower Using the Multinomial Logistic Regression

Predicted class for each flower using the multinomial logistic regression is given below:

Flower No.	Observed Species	Predicted Species	Flower No.	Observed Species	Predicted Species	Flower No.	Observed Species	Predicted Species
1	1	1	51	2	2	101	3	3
2	1	1	52	2	2	102	3	3
3	1	1	53	2	2	103	3	3
4	1	1	54	2	2	104	3	3
5	1	1	55	2	2	105	3	3
6	1	1	56	2	2	106	3	3
7	1	1	57	2	2	107	3	3
8	1	1	58	2	2	108	3	3
9	1	1	59	2	2	109	3	3
10	1	1	60	2	2	110	3	3
11	1	1	61	2	2	111	3	3
12	1	1	62	2	2	112	3	3
13	1	1	63	2	2	113	3	3
14	1	1	64	2	2	114	3	3
15	1	1	65	2	2	115	3	3
16	1	1	66	2	2	116	3	3
17	1	1	67	2	2	117	3	3
18	1	1	68	2	2	118	3	3
19	1	1	69	2	2	119	3	3
20	1	1	70	2	2	120	3	3
21	1	1	71	2	2	121	3	3
22	1	1	72	2	2	122	3	3
23	1	1	73	2	2	123	3	3
24	1	1	74	2	2	124	3	3
25	1	1	75	2	2	125	3	3
26	1	1	76	2	2	126	3	3
27	1	1	77	2	2	127	3	3
28	1	1	78	2	2	128	3	3
29	1	1	79	2	2	129	3	3
30	1	1	80	2	2	130	3	3
31	1	1	81	2	2	131	3	3
32	1	1	82	2	2	132	3	3
33	1	1	83	2	2	133	3	3
34	1	1	84	2	3	134	3	2
35	1	1	85	2	2	135	3	3
36	1	1	86	2	2	136	3	3
37	1	1	87	2	2	137	3	3
38	1	1	88	2	2	138	3	3
39	1	1	89	2	2	139	3	3
40	1	1	90	2	2	140	3	3

41	1	1	91	2	2	141	3	3
42	1	1	92	2	2	142	3	3
43	1	1	93	2	2	143	3	3
44	1	1	94	2	2	144	3	3
45	1	1	95	2	2	145	3	3
46	1	1	96	2	2	146	3	3
47	1	1	97	2	2	147	3	3
48	1	1	98	2	2	148	3	3
49	1	1	99	2	2	149	3	3
50	1	1	100	2	2	150	3	3

Classification Matrix

Predicted Response Category * Observed Response Category Crosstabulation

Count

		Observed Response Category			
		Setosa	Versicolor	Virginica	Total
Predicted Response Category	Setosa	50	0	0	50
	Versicolor	0	49	1	50
	Virginica	0	1	49	50
Total		50	50	50	150

Conclusion: Percentage of correct classification in “Setosa” is 100%.

Percentage of correct classification in “Versicolor” is 98%.

Percentage of correct classification in “Virginica” is 98%.

Multiclass Classification Vs Multinomial Classification

In this last section we will compare the 2 classification approach i.e. multiclass and multinomial classification which was performed on Flower Species data.

Comparison of the predicted class for each flower based on both the approaches:

From the predicted class table we see, in multiclass classification 3 flower species were misidentified while 2 flower species were misidentified in multinomial classification. Both classification perform equally well.

Comparison of the diagonal entries of the classification matrix based on both the approaches:

Table given below shows the diagonal entries from classification matrix on 2 approaches.

	Setosa	Versicolor	Virginica
Multiclass Classification	50	48	49
Multinomial Classification	50	49	49

Conclusion: Multinomial classification classify flower species better.

Comparison of the percentage of correct classification for both the approaches:

Table given below shows the diagonal entries from classification matrix on 2 approaches.

	Setosa	Versicolor	Virginica
Multiclass Classification	100%	96%	98%
Multinomial Classification	100%	98%	98%

Conclusion: Multinomial classification classify flower species better.