An Assignment on Classification

Course Name: Multivariate Analysis

Course Code: PM-ASDS06

Submitted to:

Md. Habibur Rahman

Associate Professor

Department of Statistics, JU

Submitted by:

Mohammad Saiduzzaman Sayed

ID: 20215063

Batch: 5th

Sec: A

Professional Masters in

Applied Statistics and Data Science (PM-ASDS)

JAHANGIRNAGAR UNIVERSITY

Assignment on Classification

Problem Statement:

Classifying Females and Males salmon

Dataset:

Salmon Data (Growth-Ring Diameters)

Description:

The salmon data frame has 100 rows and 3 columns of data from an experiment on the salmon fish.

Format:

Gender: a factor with levels *Females and Males*. Females are coded as 1 and Males are coded as 2.

Freshwater: a numeric vector of freshwater growth (hundredths of an inch).

Marine: a numeric vector of marine growth (hundredths of an inch).

Details:

Salmon fishes have a remarkable life cycle. They are born in freshwater streams and after a year or two swim into the ocean. After a couple of years in saltwater, they return to their place of birth to spawn and die. At the time they are about to return as mature fish, they are harvested while still in the ocean. To help regulate catches, samples of fish taken during the harvest must be identified based on female or male. The fish carry some information about their gender in the growth rings on their scales. Typically, the rings are associated with freshwater growth and marine growth.

Source:

Data courtesy of K.A. Jensen and B.Van Alen of the State of Alaska Department of Fish and Game.

Classification Procedure:

Here,

$X_1 =$ diameter of rings for the first-year freshwater growth (hundredths of an inch)

$X_2 =$ diameter of rings for the first-year marine growth (hundredths of an inch)

Training samples of sizes $n_1 = 52$ Females and $n_2 = 48$ males salmon yield the summary statistics,

Data matrices are,

$$\underset{(n_1 \times p)}{X_1} = \begin{bmatrix} x'_{11} \\ x'_{12} \end{bmatrix}$$

$$\underset{(n_2 \times p)}{X_2} = \begin{bmatrix} x'_{21} \\ x'_{22} \end{bmatrix}$$

the sample mean vectors and covariance matrices are determined by,

$$\underset{(p \times 1)}{\bar{x}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j}\,, \qquad \underset{(p \times p)}{S_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)\,(x_{1j} - \bar{x}_1)'$$

$$\underset{(p \times 1)}{\bar{x}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}\,, \qquad \underset{(p \times p)}{S_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)\,(x_{2j} - \bar{x}_2)'$$

The results obtained by using the above formulas:

$$\bar{x}_1 = \begin{bmatrix} 118.0577 \\ 396.3269 \end{bmatrix}, \qquad S_1 = \begin{bmatrix} 777.1143 & -642.4310 \\ -642.4310 & 1808.773 \end{bmatrix}$$

$$\bar{x}_2 = \begin{bmatrix} 117.7708 \\ 400.1042 \end{bmatrix}, \qquad S_2 = \begin{bmatrix} 580.7336 & -669.6565 \\ -669.6565 & 2533.4570 \end{bmatrix}$$

Since it is assumed that the parent populations have the same covariance matrix $\boldsymbol{\Sigma}$, the sample covariance matrices S_1 and S_2 are combined (pooled) to derive a single, unbiased estimate of $\boldsymbol{\Sigma}$. In particular, the weighted average is an unbiased estimate of $\boldsymbol{\Sigma}$ if the data matrices X_1 and X_2 contain random samples from the populations π_1 and π_2 respectively.

So, the pooled variances formula is,

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_2$$

So, the S_{pooled} is,

$$S_{pooled} = \begin{bmatrix} 682.9317 & -655.4881 \\ -655.4881 & 2156.3257 \end{bmatrix} \text{ and}$$

$$S_{pooled}^{-1} = \begin{bmatrix} 0.00207 & 0.00062 \\ 0.00062 & 0.00065 \end{bmatrix}$$

Substituting $\bar{x}_1$ for μ_1, $\bar{x}_2$ for μ_2, and S_{pooled} for $\boldsymbol{\Sigma}$ in, the "sample" classification rule:

Allocate x_0 to π_1 if

$$(\bar{x}_1 - \bar{x}_2)' \, S_{pooled}^{-1} \, x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' \, S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) \geq ln\left[\left(\frac{c(1|2)}{c(2|1)}\right) \quad \left(\frac{p_2}{p_1}\right)\right]$$

Allocate x_0 to π_2 otherwise.

If, $c(1|2) = c(2|1) \; and \; p_1 = p_2$ then,

Allocate x_0 to π_1 if

$$(\bar{x}_1 - \bar{x}_2)' \, S_{pooled}^{-1} \, x_0 \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' \, S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2)$$

Allocate x_0 to π_2 otherwise.

Where, $\hat{y} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0$ and $\hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2)$ and $\hat{a}'x = [\bar{x}_1 - \bar{x}_2]' S_{pooled}^{-1} x$

So, $\hat{y} = \hat{a}'x = [\bar{x}_1 \quad -\bar{x}_2]' S_{pooled}^{-1} x$

$$= [-0.2869 \quad 3.7773] \begin{bmatrix} 0.00207 & 0.00062 \\ 0.00062 & 0.00065 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 0.0017x_1 + 0.0022x_2$$

Moreover,

$$\bar{y}_1 = \hat{a}'\bar{x}_1 = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} \bar{x}_1$$

$$= [0.0017 \;\; 0.0022] \begin{bmatrix} 118.0577 \\ 396.3269 \end{bmatrix} = 1.0726 \text{ and}$$

$$\bar{y}_2 = \hat{a}'\bar{x}_2 = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} \bar{x}_2$$

$$= [0.0017 \;\; 0.0022] \begin{bmatrix} 117.7708 \\ 400.1042 \end{bmatrix} = 1.0804$$

So, $\hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2)$

$$= \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

$$= \frac{1}{2}(1.0726 + 1.0804) = 1.0765$$

Now if we input our first value which is $x_1 = 131 \; and \; x_2 = 355$. Should this salmon be classified as $\pi_1 \; (Female) \; or \; \pi_2 (Male)$?

We obtain,

Allocate x_0 to π_1 if $\hat{y}_0 = \hat{a}'x_0 \geq \hat{m} = 1.0765$

Allocate x_0 to π_2 if $\hat{y}_0 = \hat{a}'x_0 < \hat{m} = 1.0765$

Where $x_0' = [131, 355]$, Since

$$\hat{y}_0 = \hat{a}'x_0 = [0.0017 \;\; 0.0022] \begin{bmatrix} 131 \\ 355 \end{bmatrix} = 1.0037 < 1.0765$$

We classify this salmon fish as π_2, a male salmon fish. Now we can calculate each observation in our salmon dataset by following the steps above, and we can classify other salmon fishes as females or males.

Gender	Freshwater	Marine	$\hat{y}_0$	$\hat{m}$	lda
1	131	355	1.0037	1.0765	2
1	105	469	1.2103	1.0765	1
1	99	402	1.0527	1.0765	2
1	94	440	1.1278	1.0765	1
1	99	403	1.0549	1.0765	2
1	114	428	1.1354	1.0765	1
1	123	372	1.0275	1.0765	2
1	104	407	1.0722	1.0765	2
1	119	474	1.2451	1.0765	1
1	114	396	1.065	1.0765	2
1	109	397	1.0587	1.0765	2
1	82	431	1.0876	1.0765	1
1	105	388	1.0321	1.0765	2
1	121	403	1.0923	1.0765	1
1	85	451	1.1367	1.0765	1
1	83	453	1.1377	1.0765	1
1	53	427	1.0295	1.0765	2
1	95	411	1.0657	1.0765	2
1	76	442	1.1016	1.0765	1
1	95	426	1.0987	1.0765	1
1	70	397	0.9924	1.0765	2
1	74	451	1.118	1.0765	1
1	80	398	1.0116	1.0765	2
1	95	433	1.1141	1.0765	1
1	99	481	1.2265	1.0765	1
1	87	480	1.2039	1.0765	1
1	129	420	1.1433	1.0765	1
1	148	371	1.0678	1.0765	2
1	179	407	1.1997	1.0765	1
1	156	419	1.187	1.0765	1
1	140	362	1.0344	1.0765	2
1	108	330	0.9096	1.0765	2
1	135	355	1.0105	1.0765	2
1	152	301	0.9206	1.0765	2
1	153	397	1.1335	1.0765	1

1	152	301	0.9206	1.0765	2
1	148	383	1.0942	1.0765	1
1	145	337	0.9879	1.0765	2
1	123	364	1.0099	1.0765	2
1	117	355	0.9799	1.0765	2
1	118	379	1.0344	1.0765	2
1	153	403	1.1467	1.0765	1
1	154	390	1.1198	1.0765	1
1	155	349	1.0313	1.0765	2
1	128	400	1.0976	1.0765	1
1	144	403	1.1314	1.0765	1
1	133	375	1.0511	1.0765	2
1	128	383	1.0602	1.0765	2
1	144	373	1.0654	1.0765	2
1	125	346	0.9737	1.0765	2
1	153	352	1.0345	1.0765	2
1	108	339	0.9294	1.0765	2
2	108	368	0.9932	1.0765	2
2	86	506	1.2594	1.0765	1
2	87	423	1.0785	1.0765	1
2	117	489	1.2747	1.0765	1
2	79	432	1.0847	1.0765	1
2	123	372	1.0275	1.0765	2
2	109	420	1.1093	1.0765	1
2	112	394	1.0572	1.0765	2
2	111	422	1.1171	1.0765	1
2	126	423	1.1448	1.0765	1
2	105	434	1.1333	1.0765	1
2	100	470	1.204	1.0765	1
2	84	399	1.0206	1.0765	2
2	102	429	1.1172	1.0765	1
2	101	469	1.2035	1.0765	1
2	85	444	1.1213	1.0765	1
2	106	442	1.1526	1.0765	1
2	118	381	1.0388	1.0765	2
2	87	402	1.0323	1.0765	2
2	84	511	1.267	1.0765	1
2	91	469	1.1865	1.0765	1
2	101	474	1.2145	1.0765	1
2	92	404	1.0452	1.0765	2
2	94	491	1.24	1.0765	1
2	152	381	1.0966	1.0765	1
2	166	377	1.1116	1.0765	1

2	124	389	1.0666	1.0765	2
2	131	345	0.9817	1.0765	2
2	144	345	1.0038	1.0765	2
2	149	393	1.1179	1.0765	1
2	170	386	1.1382	1.0765	1
2	136	438	1.1948	1.0765	1
2	122	306	0.8806	1.0765	2
2	90	385	1	1.0765	2
2	145	376	1.0737	1.0765	2
2	115	354	0.9743	1.0765	2
2	134	383	1.0704	1.0765	2
2	126	345	0.9732	1.0765	2
2	120	369	1.0158	1.0765	2
2	150	354	1.0338	1.0765	2
2	109	325	0.9003	1.0765	2
2	117	344	0.9557	1.0765	2
2	163	370	1.0911	1.0765	1
2	145	355	1.0275	1.0765	2
2	123	349	0.9769	1.0765	2
2	140	388	1.0916	1.0765	1
2	150	339	1.0008	1.0765	2
2	124	341	0.961	1.0765	2

Prior probabilities of groups:

1 (Female)	2 (Male)
0.52	0.48

Counts:

1 (Female)	2 (Male)
52	48

Coefficients of linear discriminants:

	LD
Freshwater	0.01972598
Marine	0.02539926

Confusion Matrices:

		Predicted	
		1 (Female)	2 (Male)
Actual	1 (Female)	23	29
	2 (Male)	24	24

Accuracy rate:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$= \frac{23 + 24}{23 + 24 + 29 + 24} = 0.47$$

So, the accuracy comes out to 0.47 or 47%. That means our model is not good of identifying genders.

THE END