University of Glasgow | School of Computing Science

CSC3005 Laboratory/Tutorial 2 Solution: Data Preprocessing and Classification Analysis I

5. **Theory on PCA**

Step 1: Find covariance matrix

Data matrix $\mathbf{X} = \begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix}$

$$\mathrm{cov}(\mathbf{X}) = \frac{1}{N-1}(\mathbf{X}-\boldsymbol{\mu})^T(\mathbf{X}-\boldsymbol{\mu})$$

where $\boldsymbol{\mu}$ is the mean of the data for each column and N is the number of data record.

$$\boldsymbol{\mu} = \begin{bmatrix} \frac{4+0}{2} & \frac{0+8}{2} \\ \frac{4+0}{2} & \frac{0+8}{2} \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$$

Therefore

$$\mathrm{cov}(\mathbf{X}) = \frac{1}{2-1}\left(\begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix} - \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}\right)^T \left(\begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix} - \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}\right)$$

$$\mathrm{cov}(\mathbf{X}) = \left(\begin{bmatrix} 2 & -4 \\ -2 & 4 \end{bmatrix}\right)^T \left(\begin{bmatrix} 2 & -4 \\ -2 & 4 \end{bmatrix}\right) = \begin{bmatrix} 2 & -2 \\ -4 & 4 \end{bmatrix}\begin{bmatrix} 2 & -4 \\ -2 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 2*2+(-2*-2) & 2*-4+(-2*4) \\ -4*2+(4*-2) & -4*-4+4*4 \end{bmatrix}$$

$$= \begin{bmatrix} 8 & -16 \\ -16 & 32 \end{bmatrix}$$

Step 2: Find eigenvalues and eigenvectors

Let λ be the eigenvalues of the $\mathrm{cov}(\mathbf{X})$

$$\det(\lambda\mathbf{I}\text{-}\mathrm{cov}(\mathbf{X}))=0$$

For $\lambda\mathbf{I}\text{-}\mathrm{cov}(\mathbf{X})$

$$\lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 8 & -16 \\ -16 & 32 \end{bmatrix}$$

$$\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 8 & -16 \\ -16 & 32 \end{bmatrix} = \begin{bmatrix} \lambda\text{-}8 & 16 \\ 16 & \lambda\text{-}32 \end{bmatrix}$$

Therefore

$$\det(\lambda\mathbf{I}\text{-}\mathrm{cov}(\mathbf{X}))=0$$

$$\left|\begin{bmatrix} \lambda\text{-}8 & 16 \\ 16 & \lambda\text{-}32 \end{bmatrix}\right| = 0$$
$$(\lambda\text{-}8)(\lambda\text{-}32) - 16 * 16 = 0$$
$$\lambda^2 - 32\lambda - 8\lambda + 256 - 256 = 0$$
$$\lambda^2 - 40\lambda = 0$$
$$\lambda(\lambda - 40) = 0 \rightarrow \lambda = 0 \; or \; \lambda = 40$$

For $\lambda = 0$

$$(\lambda\mathbf{I}\text{-cov}(\mathbf{X})) \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$
$$\begin{bmatrix} \lambda\text{-}8 & 16 \\ 16 & \lambda\text{-}32 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$
$$\begin{bmatrix} 0\text{-}8 & 16 \\ 16 & 0\text{-}32 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$
$$\begin{bmatrix} \text{-}8 & 16 \\ 16 & \text{-}32 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

Using first row

$$-8v_1 + 16v_2 = 0$$
$$\rightarrow \quad v_1 = 2v_2$$

Therefore first eigenvector is

$$\mathbf{v_1} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

if $v_2 = 1$

For $\lambda = 40$

$$(\lambda\mathbf{I}\text{-cov}(\mathbf{X})) \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} \lambda\text{-}8 & 16 \\ 16 & \lambda\text{-}32 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$
$$\begin{bmatrix} 40\text{-}8 & 16 \\ 16 & 40\text{-}32 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$
$$\begin{bmatrix} 32 & 16 \\ 16 & 8 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

Using first row

$$32v_1 + 16v_2 = 0$$
$$\rightarrow \quad v_1 = -0.5v_2$$

Therefore second eigenvector is

$$\mathbf{v_2} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 1 \end{bmatrix}$$

if $v_2 = 1$

Since second eigenvector has the eigenvalue of 40 which is the largest power, therefore the second eigenvector will be used as the primary eigenvector principal component

The projection of PC1 on data will be given as

$$(\mathbf{X}-\boldsymbol{\mu}).\mathbf{v_2} = \left(\begin{bmatrix}4 & 0\\0 & 8\end{bmatrix} - \begin{bmatrix}2 & 4\\2 & 4\end{bmatrix}\right).\begin{bmatrix}-0.5\\1\end{bmatrix}$$
$$= \begin{bmatrix}2 & -4\\-2 & 4\end{bmatrix}\begin{bmatrix}-0.5\\1\end{bmatrix}$$
$$= \begin{bmatrix}2 * -0.5 - 4\\-2 * -0.5 + 4\end{bmatrix} = \begin{bmatrix}-5\\5\end{bmatrix}$$

Hence 2x2 matrix of $\mathbf{X}$ that contains 2 features become 2x1 vectors that contains one PC component that capture the 2 features. Hence, dimension reduction!!!

7. Work out the Impurity Measure and Information Gain for Decision Tree

Step 1: Work out the Parent Entropy for the two classes

	Name	Warm-blooded	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class
0	human	1	1	0	0	1	0	mammals
1	python	0	0	0	0	0	1	non-mammals
2	salmon	0	0	1	0	0	0	non-mammals
3	whale	1	1	1	0	0	0	mammals
4	frog	0	0	1	0	1	1	non-mammals
5	komodo	0	0	0	0	1	0	non-mammals
6	bat	1	1	0	1	1	1	mammals
7	pigeon	1	0	0	1	1	0	non-mammals
8	cat	1	1	0	0	1	0	mammals
9	leopard shark	0	1	1	0	0	0	non-mammals
10	turtle	0	0	1	0	1	0	non-mammals
11	penguin	1	0	1	0	1	0	non-mammals
12	porcupine	1	1	0	0	1	1	mammals
13	eel	0	0	1	0	0	0	non-mammals
14	salamander	0	0	1	0	1	1	non-mammals

Parent (root)	Count	Probability of Class
Class 1 : Mammals	5	$p_1 = \frac{5}{15}$
Class 2 : Non Mammals	10	$p_2 = \frac{10}{15}$
Entropy =- $\sum_{i=1}^{2} p_i \, log_2(pi) = -p_1 log_2 p_1 - p_2 log_2 p_2 = -\frac{5}{15} log_2 \frac{5}{15} - \frac{10}{15} log_2 \frac{10}{15}$ =**0.918**		

Step 2: Find the best split. Work out the Entropy and Information Gain for all features by splitting all feature into two Node.

Warm Blooded	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	5	$p_1 = \frac{5}{7}$	0	$p_1 = \frac{0}{8}$
Class 2: Non-Mammals	2	$p_2 = \frac{2}{7}$	8	$p_2 = \frac{8}{8}$
$Entropy$		$-\frac{5}{7} log_2 \frac{5}{7} - \frac{2}{7} log_2 \frac{2}{7} = 0.863$		$-\frac{0}{8} log_2 \frac{0}{8} - \frac{8}{8} log_2 \frac{8}{8} = 0$
$Entropy_{Split}$	$\frac{7}{15} Entropy_{N1} + \frac{8}{15} Entropy_{N2} = \frac{7}{15} * 0.863 + \frac{8}{15} * 0 = 0.403$			
Δ_{info}	0.918-0.403=0.515			

Give Birth	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	5	$p_1 = \frac{5}{6}$	0	$p_1 = \frac{0}{9}$
Class 2: Non-Mammals	1	$p_2 = \frac{1}{6}$	9	$p_2 = \frac{9}{9}$
$Entropy$		$-\frac{5}{6} log_2 \frac{5}{6} - \frac{1}{6} log_2 \frac{1}{6} = 0.65$		$-\frac{0}{9} log_2 \frac{0}{9} - \frac{9}{9} log_2 \frac{9}{9} = 0$
$Entropy_{Split}$	$\frac{6}{15} Entropy_{N1} + \frac{9}{15} Entropy_{N2} = \frac{6}{15} * 0.65 + \frac{9}{15} * 0 = 0.26$			
Δ_{info}	0.918-0.26=**0.658**			

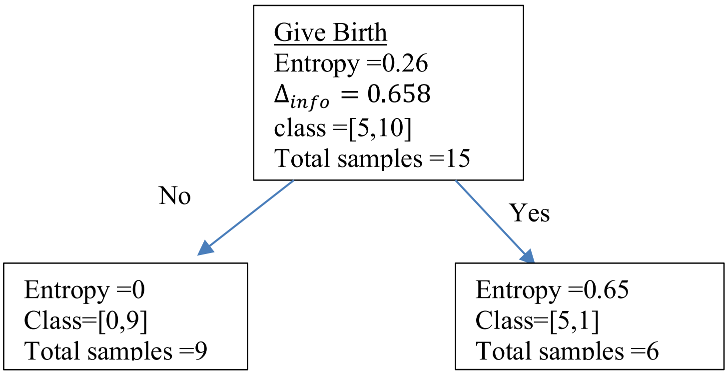
Aquatic Creature	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	1	$p_1 = \frac{1}{8}$	4	$p_1 = \frac{4}{7}$
Class 2: Non-Mammals	7	$p_2 = \frac{7}{8}$	3	$p_2 = \frac{3}{7}$
$Entropy$		$-\frac{1}{8}log_2\frac{1}{8} - \frac{7}{8}log_2\frac{7}{8} = 0.544$		$-\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7} = 0.985$
$Entropy_{Split}$	$\frac{8}{15}Entropy_{N1} + \frac{7}{15}Entropy_{N2} = \frac{8}{15} * 0.544 + \frac{7}{15} * 0.985 = 0.750$			
Δ_{info}	0.918-0.750=0.169			

Aerial Creature	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	1	$p_1 = \frac{1}{2}$	4	$p_1 = \frac{4}{13}$
Class 2: Non-Mammals	1	$p_2 = \frac{1}{2}$	9	$p_2 = \frac{9}{13}$
$Entropy$		$-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1$		$-\frac{4}{13}log_2\frac{4}{13} - \frac{9}{13}log_2\frac{9}{13} = 0.89$
$Entropy_{Split}$	$\frac{2}{15}Entropy_{N1} + \frac{13}{15}Entropy_{N2} = \frac{2}{15} * 1 + \frac{13}{15} * 0.89 = 0.905$			
Δ_{info}	0.918-0.905=0.013			

Has Legs	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	4	$p_1 = \frac{4}{10}$	1	$p_1 = \frac{1}{5}$
Class 2: Non-Mammals	6	$p_2 = \frac{6}{10}$	4	$p_2 = \frac{4}{5}$
$Entropy$		$-\frac{4}{10}log_2\frac{4}{10} - \frac{6}{10}log_2\frac{6}{10} = 0.971$		$-\frac{1}{5}log_2\frac{1}{5} - \frac{4}{5}log_2\frac{4}{5} = 0.722$
$Entropy_{Split}$	$\frac{10}{15}Entropy_{N1} + \frac{5}{15}Entropy_{N2} = \frac{10}{15} * 0.971 + \frac{5}{15} * 0.722 = 0.888$			
Δ_{info}	0.918-0.888=0.003			

Hibernates	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	2	$p_1 = \frac{2}{5}$	3	$p_1 = \frac{3}{10}$
Class 2: Non-Mammals	3	$p_2 = \frac{3}{5}$	7	$p_2 = \frac{7}{10}$
$Entropy$		$-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$		$-\frac{3}{10}log_2\frac{3}{10} - \frac{7}{10}log_2\frac{7}{10} = 0.881$
$Entropy_{Split}$	$\frac{5}{15}Entropy_{N1} + \frac{10}{15}Entropy_{N2} = \frac{5}{15} * 0.971 + \frac{10}{15} * 0.881 = 0.9111$			
Δ_{info}	0.918-0.9111=0.007			

Give Birth Feature has the highest information gain of **0.658**, therefore it will be selected as the first node. Under Give Birth Feature, N2 node has the pure classification with [mammal, non-mammal]=[0 9]. However, its N1 node has impurity in classification with [mammal, non-mammal]=[5 1]. There is 1 data that results 1 label of non-mammal. As such, this N1 node need to split further.



Step 3 : Repeat the whole process as in step 2 but with Birth feature N1 node as the parent node that has entropy =0.65 and Give Birth=1

Warm Blooded	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	5	$p_1 = \frac{5}{5}$	0	$p_1 = \frac{0}{1}$
Class 2: Non-Mammals	0	$p_2 = \frac{0}{5}$	1	$p_2 = \frac{1}{1}$
$Entropy$		$-\frac{5}{5}log_2\frac{5}{5} - \frac{0}{5}log_2\frac{0}{5} = 0$		$-\frac{0}{1}log_2\frac{0}{1} - \frac{1}{1}log_2\frac{1}{1} = 0$
$Entropy_{Split}$	$\frac{5}{6}Entropy_{N1} + \frac{1}{6}Entropy_{N2} = \frac{5}{6} * 0 + \frac{1}{6} * 0 = 0$			
Δ_{info}	0.65-0=**0.65**			

Aquatic Creature	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	1	$p_1 = \frac{1}{2}$	4	$p_1 = \frac{4}{4}$
Class 2: Non-Mammals	1	$p_2 = \frac{1}{2}$	0	$p_2 = \frac{0}{4}$
$Entropy$		$-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1$		$-\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$
$Entropy_{Split}$	$\frac{2}{6}Entropy_{N1} + \frac{4}{6}Entropy_{N2} = \frac{2}{6} * 1 + \frac{4}{6} * 0 = 0.333$			
Δ_{info}	0.65-0.333=0.317			

Aerial Creature	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	1	$p_1 = \frac{1}{1}$	4	$p_1 = \frac{4}{5}$
Class 2: Non-Mammals	0	$p_2 = \frac{0}{1}$	1	$p_2 = \frac{1}{5}$
$Entropy$		$-\frac{1}{1}log_2\frac{1}{1} - \frac{0}{1}log_2\frac{0}{1} = 0$		$-\frac{4}{5}log_2\frac{4}{5} - \frac{1}{5}log_2\frac{1}{5} = 0.722$
$Entropy_{Split}$	$\frac{1}{6}Entropy_{N1} + \frac{5}{6}Entropy_{N2} = \frac{1}{6} * 0 + \frac{5}{6} * 0.722 = 0.602$			
Δ_{info}	0.65-0.602=0.048			

Has Legs	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	4	$p_1 = \frac{4}{4}$	1	$p_1 = \frac{1}{2}$
Class 2: Non-Mammals	0	$p_2 = \frac{0}{4}$	1	$p_2 = \frac{1}{2}$
$Entropy$		$-\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$		$-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1$
$Entropy_{Split}$	$\frac{4}{6}Entropy_{N1} + \frac{2}{6}Entropy_{N2} = \frac{4}{6} * 0 + \frac{2}{6} * 1 = 0.333$			
Δ_{info}	0.65-0.333=0.317			

Hibernates	N1 (Yes=1)	N1 Probability of Class	N2 (No=0)	N2 Probability of Class
Class 1: Mammals	2	$p_1 = \frac{2}{2}$	3	$p_1 = \frac{3}{4}$
Class 2: Non-Mammals	0	$p_2 = \frac{0}{2}$	1	$p_2 = \frac{1}{4}$
$Entropy$		$-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2} = 0$		$-\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} = 0.811$

$Entropy_{Split}$	$\frac{2}{6}Entropy_{N1} + \frac{4}{6}Entropy_{N2} = \frac{2}{6} * 0 + \frac{4}{6} * 0.811 = 0.541$
Δ_{info}	0.65-0.541=0.109

Warm Blooded Feature has the highest information gain of **<u>0.65</u>**, therefore it will be selected as the second node. Under warm blooded Feature, both N1 and N2 node has the pure classification with [mammal, non-mammal]=[5 0] and [0 1] respectively. As such, no need to split further. The Decision Tree Classifier is the same design as in Question 6 solution