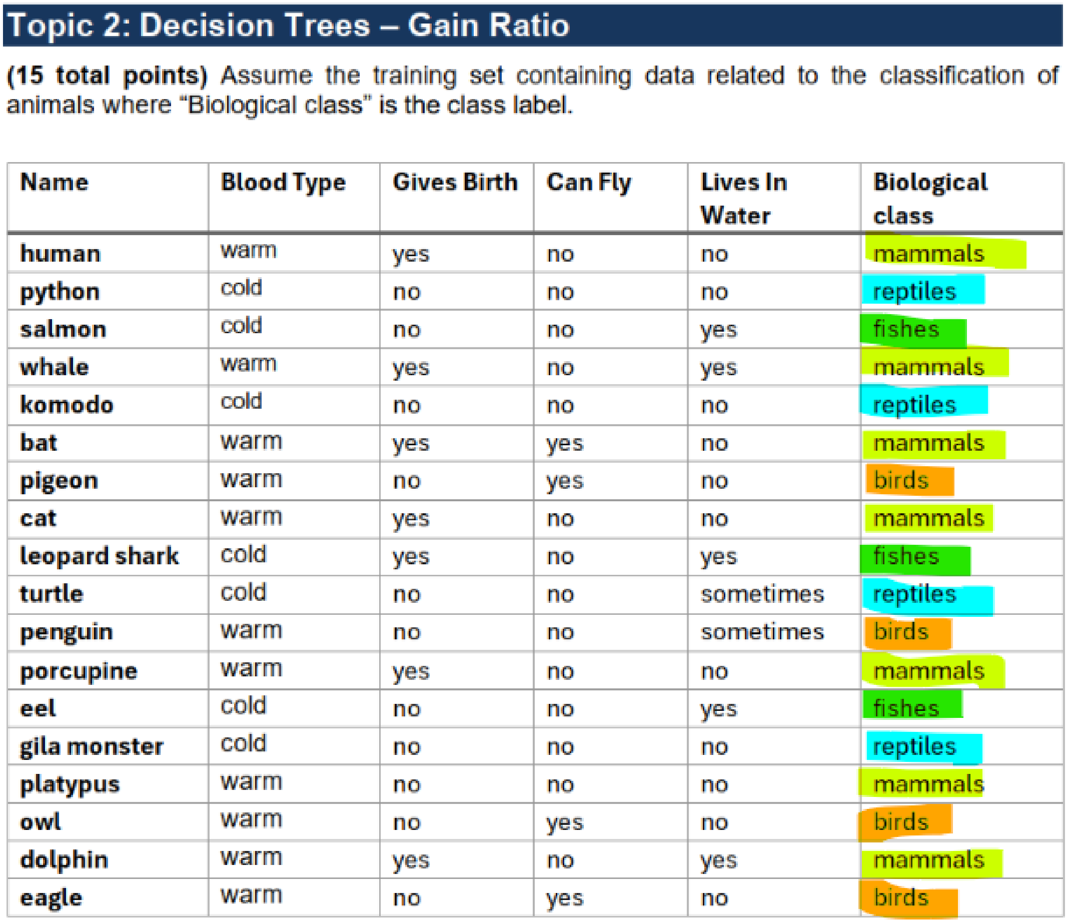
Panagiotis Paltsokas std163861@ac.eap.gr

DAMA 60

Written Assignment 1 extra material

Topic 2: Decision Trees – Gain Ratio



Topic 2: Decision Trees – Gain Ratio

(15 total points) Assume the training set containing data related to the classification of animals where "Biological class" is the class label.

Name	Blood Type	Gives Birth	Can Fly	Lives In Water	Biological class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

The distribution of the class label is as follows :

Mammals = 7/18

Reptiles = 4/18

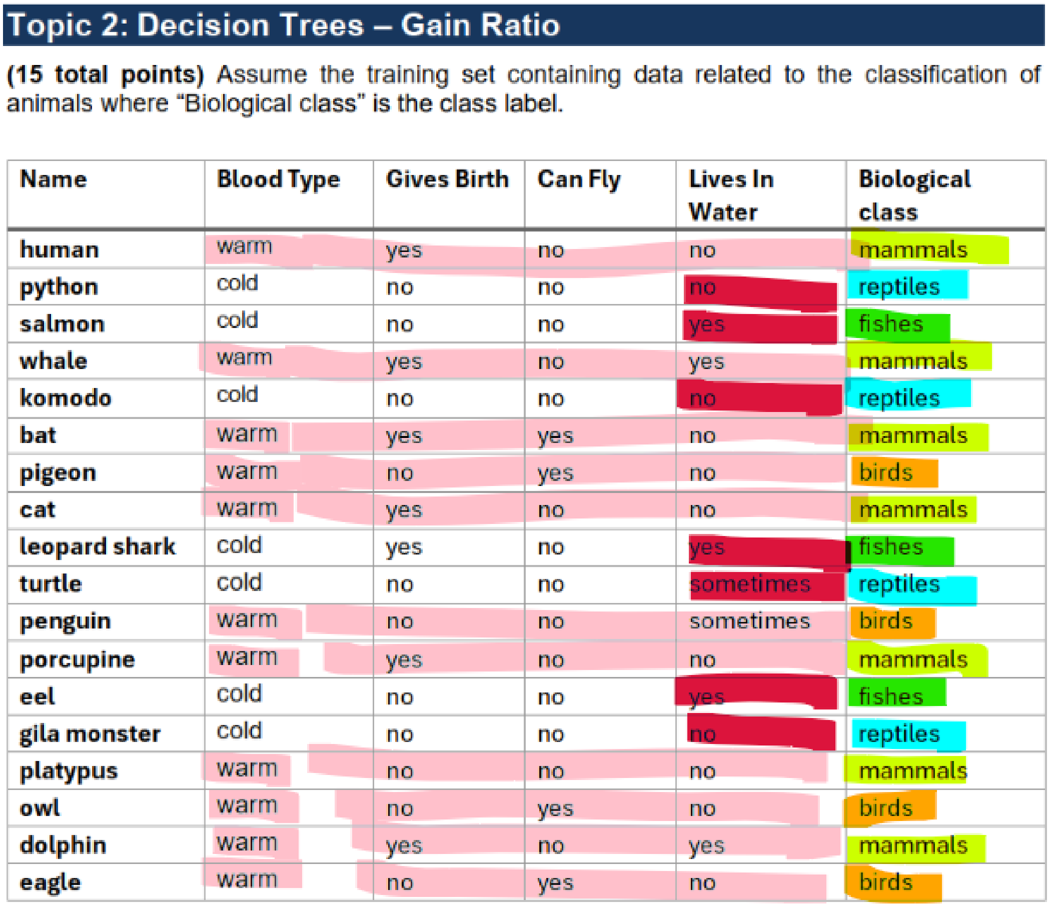
Fishes = 3/18

Birds = 4/18

We calculate the Entropy before the split which is:

$$\text{Entropy}_{\text{Before}} = \text{E}\left(\frac{7}{18},\frac{2}{9},\frac{1}{6},\frac{2}{9}\right) = -\left(\frac{7}{18}\log_2\frac{7}{18} + 2\cdot\frac{2}{9}\log_2\frac{2}{9} + \frac{1}{6}\log_2\frac{1}{6}\right) = ... = 1.9251$$

Split Feature : Blood Type



Topic 2: Decision Trees – Gain Ratio

(15 total points) Assume the training set containing data related to the classification of animals where "Biological class" is the class label.

Name	Blood Type	Gives Birth	Can Fly	Lives In Water	Biological class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

Warm & Mammals = 7/11 Warm & birds = 4/11 Cold & reptiles = 4/7 Cold & fishes = 3/7

$$\text{Entropy}_{\text{Warm}} = E\left(\frac{7}{11},\frac{4}{11}\right) = ... = 0.9456 \qquad \text{Entropy}_{\text{Cold}} = E\left(\frac{4}{7},\frac{3}{7}\right) = ... = 0.9852$$

$$\text{weight}_{\text{Warm}} = \frac{11}{18} \qquad \text{weight}_{\text{Cold}} = \frac{7}{18}$$

$$\text{SplitInfo} = -\frac{11}{18}\log_2\frac{11}{18} - \frac{7}{18}\log_2\frac{7}{18} = ... = 0.9640$$

$$\text{Entropy}_{\text{After}} = \text{Weight}_{\text{Warm}}\cdot\text{Entropy}_{\text{Warm}} + \text{Weight}_{\text{Cold}}\cdot\text{Entropy}_{\text{Cold}} = \frac{11}{18}\cdot 0.9456 + \frac{7}{18}\cdot 0.9852 = 0.9610$$

$$\text{Information Gain} = \text{Entropy}_{\text{Before}} - \text{Entropy}_{\text{After}} = 0.9640$$

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{SplitInfo}} = \frac{0.9640}{0.9640} = 1$$

For the features "Gives Birth", "Can Fly", and "Lives In Water", we work in the same manner.

Topic 4: Clustering – K-means/Hierarchical

a)

We calculate the missing distances $d\left(P_{i,}P_j\right)=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2+\left(z_i-z_j\right)^2}$

for $i,j=0,1,2,\ldots$ and points $P_i=(x_i,y_i,z_i)$, and $P_j=\left(x_{j,}y_j,z_j\right)$

$$d\left(P_1,P_5\right)=\sqrt{\left(0.1\right)^2+\left(0.1\right)^2+\left(0.2\right)^2}=\sqrt{0.06}=0.2449$$

$$d\left(P_2,P_4\right)=\sqrt{\left(0.1\right)^2+\left(0.1\right)^2+\left(0.2\right)^2}=\sqrt{0.06}=0.2449$$

$$d\left(P_3,P_6\right)=\sqrt{0+\left(0.1\right)^2+\left(0.4\right)^2}=\sqrt{0.17}=0.4123$$

$$d\left(P_4,P_5\right)=\sqrt{\left(0.4\right)^2+\left(0.2\right)^2+\left(0.6\right)^2}=\sqrt{0.56}=0.7483$$

$$d\left(P_5,P_6\right)=\sqrt{\left(0.6\right)^2+\left(0.2\right)^2+0}=\sqrt{0.4}=0.6324$$

b) Initial Centers : $P_1\left(0.4,\ 1.2,\ 1.7\right)$, $P_2\left(0.6,\ 0.8,\ 1.1\right)$

$$d\left(P_3,P_1\right)=\ldots=0.5$$

$$d\left(P_3,P_2\right)=\ldots=0.6708$$

$$d\left(P_4,P_1\right)=\ldots=0.9055$$

$$d\left(P_4,P_2\right)=\ldots=0.2449$$

$$d\left(P_5,P_1\right)=\ldots=0.2449$$

$$d\left(P_5,P_2\right)=\ldots=0.5830$$

$$d\left(P_6,P_1\right)=\ldots=0.8602$$

$$d\left(P_6,P_2\right)=\ldots=0.6164$$

Choosing the smallest distance from the initial centers, we create the clusters $C_1=P_1,\ P_3,\ P_5$ and $C_2=P_2,\ P_4,\ P_6$

We calculate the new centroids' coordinates:

$$x_{C_1}=\frac{x_1+x_3+x_5}{3}=\frac{0.4+0.1+0.3}{3}=\frac{0.8}{3}=0.2666$$

$$y_{C_1}=\frac{y_1+y_3+y_5}{3}=\frac{1.2+1.2+1.1}{3}=\frac{3.5}{3}=1.1666$$

$$z_{C_1}=\frac{z_1+z_3+z_5}{3}=\frac{1.7+1.3+1.5}{3}=\frac{4.5}{3}=1.5$$

$$x_{C_2}=\frac{x_2+x_4+x_6}{3}=....=0.4666$$

$$y_{C_2}=\frac{y_2+y_4+y_6}{3}=\ldots=0.9333$$

$$z_{C_2}=\frac{z_2+z_4+z_6}{3}=\ldots=0.9666$$

The new cluster centers, after the first iteration of the K-means algorithm, and before starting the second, are:

$C_1 = (0.2666,\ 1.1666,\ 1.5)$ and $C_2 = (0.4666,\ 0.9333,\ 0.9666)$

c)

i) For the initial pairing, the algorithm pairs together c1: {p1, p5} and c2: {p2, p4} for having the smallest distance. Then, Complete Linkage (MAX) calculates the distance between the clusters, using the maximum distance between any two points in those clusters. Therefore,

$d(c_1, c_2) = \max\{d(p_1, p_2),\ d(p_1, p_4),\ d(p_5, p_2),\ d(p_5, p_4)\} = \max\{0.7483, 0.9055, 0.5830, 0.7483\} = 0.9055$

We repeat for the clusters $c_2 = (p_2, p_4)$ and $c_3 = p_6$.

$d(c_1, c_3) = \max\{d(p_2, p_6),\ d(p_4, p_6)\} = \max\{0.6164, 0.6324\} = 0.6324$