WORKSHEET 3

PROBLEM 1 : NOTES

K- NEAREST NEIGHBOURS

Note : $\mathcal{L}_p$ - norm $= \left(\sum_{j=1}^{p} |x_{nj}|^p \right)^{1/p}$

The Euclidean norm is the $\mathcal{L}_2$ - norm.

The Euclidean distance is defined as

$$\partial(\underline{x}_i, \underline{x}_j) = \sqrt{\sum_{d=1}^{p} (x_{id} - x_{jd})^2}$$

for $\underline{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $x_{ij} \in \mathbb{R}$

$\underline{x}_j = (x_{j1}, \ldots, x_{jp})^T$, $x_{jd} \in \mathbb{R}$

Classification :

Given $\{t_i, \underline{x}_i\}_{i=1}^{N}$ $\quad t_i \in \{\mathcal{C}_1, \ldots, \mathcal{C}_C\}$

That is, the target t_i belongs to one of C classes.

Then in order to classify some unknown target observed at the location $\underline{x}_*$,

first identify the k observeations for which $\partial(\underline{x}_*, \underline{x}_i)$ is smallest to find the subset $\{t_j, \underline{x}_j\}_{j=1}^{k}$

Then $t_* = \underset{u \in \{z_1, \dots, z_L\}}{\arg\max} \sum_{j=1}^{K} \delta(u, t_j)$

where $\delta(a,b) = \begin{cases} 1 & , \text{if } a = b \\ 0 & , \text{otherwise} \end{cases}$

is a delta function

Which is to say that the predicted value / class for t_* is the class to which most of it's neighbours belong.

Classification error is of the form

$$1 - \delta(t_n, t_n^*)$$

where t_n is the "true" value of the prediction t_n^*.

PROBLEM 2

FIND THE ID3 CLASSIFIER FOR THE PROVIDED DATASET

NOTE: A GENERAL DESCRIPTION OF THE ID3 ALGORITHM.

GIVEN A DATA SET $S = \{ t_n, \underline{x}_n \}_{n=1}^{N}$

where $t_n \in \{ z_1, \dots, z_C \}$ is the target variable which belongs to one of C classes

and $\underline{x}_n = (x_{n1}, \dots, x_{nD})^T$ is a vector of explanatory variables where $x_{nj} \in Y_j$, that is to say that the j^{th} explanatory variable comes from some set Y_j.

when $p_i = \frac{1}{N} \sum_{n=1}^{N} \delta(t_n, z_i)$,

δ being a delta function (see problem 1 notes)

The entropy of S is

$$\mathcal{E}(S) = - \sum_{i=1}^{C} p_i \log_2(p_i)$$

$\mathcal{E}(S)$ is a measure of the uncertainty of the dataset where $\mathcal{E}(S) \in [0, 1]$ and

for $\mathcal{E}(S) = 0$, $C = 1$

and $\mathcal{E}(S) = 1$, $p_i = \frac{1}{C}$

ID3 provides a classification tree by splitting the data by the unused attribute which gives the greatest information gain

For each attribute, denoted A, or more precisely $x_{\cdot j} = (x_{1j}, \ldots, x_{Nj})$

The Information Gain is given by

$$IG(S, A) = IG(S, x_{\cdot j}) = \mathcal{E}(S) - \sum_{y \in Y_j} \frac{|S_{[x_{\cdot j} = y]}|}{|S|} \mathcal{E}(S_{[x_{\cdot j} = y]})$$

where $S_{[x_{\cdot j} = y]}$ is the set of observations in S for which $x_{\cdot j} = y$ and $|S|$ is the number of observations in S

Applying the algorithm to the provided dataset:

$\mathcal{E}(S)$ where $t = Y$, $x_{\cdot 1} = V$, $x_{\cdot 2} = W$, $x_{\cdot 3} = X$

$$\mathcal{E}(S) = -\left(\tfrac{2}{5}\log_2\left(\tfrac{2}{5}\right) + \tfrac{3}{5}\log_2\tfrac{3}{5}\right)$$

$$\approx 0.97$$

$$\mathcal{E}\left(S_{[X=1]}\right) = -\left(0 \cdot \log_2(0) + 1 \cdot \log_2(1)\right) = 0$$

$$\mathcal{E}\left(S_{[X=0]}\right) = -\left(\tfrac{1}{2}\log_2\left(\tfrac{1}{2}\right) + \tfrac{1}{2}\log_2\left(\tfrac{1}{2}\right)\right) = 1$$

$$\mathcal{E}(S_{[W=1]}) = \mathcal{E}(S_{[V=1]}) = -\left(\tfrac{1}{3}\log_2(\tfrac{1}{3}) + \tfrac{2}{3}\log_2(\tfrac{2}{3})\right)$$

$$\approx 0.91$$

$$\mathcal{E}(S_{[W=0]}) = \mathcal{E}(S_{[V=0]}) = 1$$

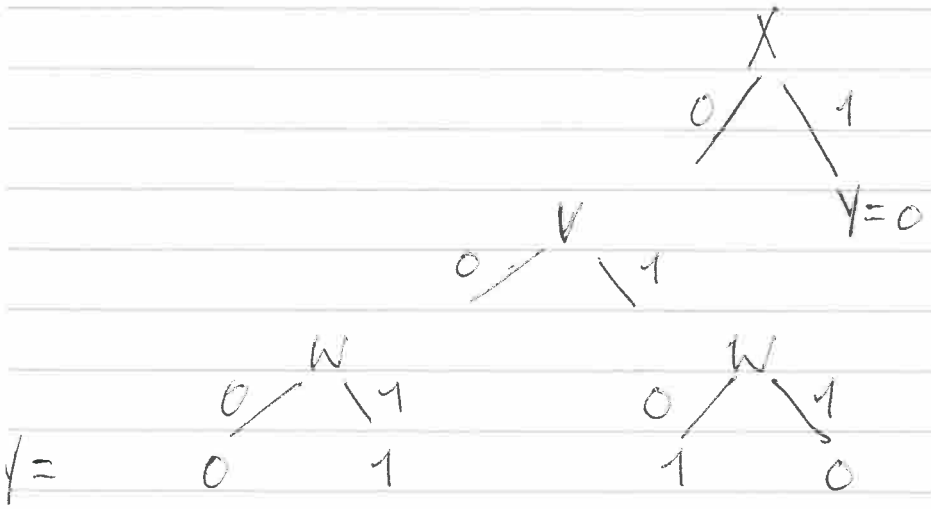
$$IG(S,X) \approx 0.97 - \tfrac{3}{5}(0) - \tfrac{4}{5}(1) \approx 0.17$$

$$IG(S,W) = IG(S,V) \approx 0.97 - \tfrac{3}{5}(1) - \tfrac{3}{5}(0.91) \approx 0.02$$

Thus the first split in the tree is on attribute X.

Continuing the process for the next split in the tree will show W and V to be equally explanatory i.e $IG(S_{[X=0]}, W) = IG(S_{[X=0]}, V)$

a ~~the~~ final tree produced by the algorithm is (there are two possible solutions)



And Y can be fully classified by X, W, V where W, and V are interchangeable