IS5152 Data-driven Decision Making
SEMESTER II 2022-2023
Assignment 4 - Suggested solution

1. (10 points) The table below shows training data from a credit scoring database.

University Education	Income Level	Profession	Status
No	High	Self-employed	Bad
Yes	Medium	White-collar	Good
No	Medium	White-collar	Bad
Yes	High	Blue-collar	Good
Yes	Low	Self-employed	Bad

Status is the class label (target attribute), while *University Education*, *Income Level* and *Profession* are the input attributes. We are building a binary classification tree using GINI index to measure impurity of the data.

(a) (2 points) What is the GINI index for the observations in this data set?

$p_{BAD} = 3/5, p_{GOOD} = 2/5$**, Gini** $= 1 - (3/5)^2 - (2/5)^2 = 12/25$**.**

(b) (4 points) List all possible splits that must be considered at the root node.

- University Education: one possible split, Yes vs No.
- Income Level: two possible splits, Low vs (Medium,High) and (Low,Medium) vs High.
- Profession: three possible splits, Self-Employed vs (White-collar, Blue-Collar), White-Collar vs (Self-Employed, Blue-Collar), and Blue-Collar vs (Self-Employed, White-Collar).

In total there are 6 possible splits.

(c) (4 points) Which of these two attributes: {University Education, Income Level} will <u>not</u> be selected as the attribute for splitting the root node? Explain your answer.

- University Education:
  - No: 2 Bad, Gini = 0.
  - Yes: 1 Bad, 2 Good, Gini= $1 - (1/3)^2 - (2/3)^2 = 4/9$.
  - Gini index after split= $0 + (3/5) \times (4/9) = 4/15 = 0.267$

- Income Level:
  - Split 1:
    * Low: 1 Bad, Gini = 0
    * Medium or High: 2 Bad, 2 Good, Gini $=1-(1/2)^2-(1/2)^2=1/2$
    * Gini index after split $= 0+(4/5)\times(1/2)=2/5=0.4$
  - Split 2:
    * Low or Medium: 2 Bad, 1 Good, Gini $= 4/9$
    * High: 1 Bad, 1 Good, Gini $= 1-(1/2)^2-(1/2)^2=1/2$.
    * Gini index after split $= (3/5)\times(4/9)+(2/5)\times(1/2)=7/15=0.467$

  The two possible splits using Income Level do not produce GINI index that is better than University Education. Hence, the first split is University Education or possibly, Profession.

2. (10 points) A system analyst studied the effect of computer programming experience on ability to complete within specified time a complex programming task. Ten persons who had varying amount of experience (in months) were selected for the study. All persons were given the same programming task, and the results of their success in the task are shown in the table below. The results are coded in binary fashion: $d = 1$ if the task was completed successfully in the alloted time, $d = 0$ otherwise. We are interested in building a model to predict if the given task can be completed successfully within the alloted time using the amount of experience (in months) as the only input.

Person	1	2	3	4	5	6	7	8	9	10
Experience (months)	14	15	6	8	29	10	25	12	30	18
Task success	0	0	0	0	1	1	1	1	1	1

For the questions below, consider only binary decision trees, that is, any node in the trees can have at most two branches.

(a) (2 points) What is the entropy of the training samples with respect to the classification?

First sort the data according to the attribute "Experience":

Person	1	2	3	4	5	6	7	8	9	10
Experience (months)	6	8	10	12	14	15	18	25	29	30
Task success	0	0	1	1	0	0	1	1	1	1

$Entropy(S) = -p_1\log_2 p_1 - p_2\log_2 p_2 = -0.4\log_2 0.4 - 0.6\log_2 0.6 = 0.97095.$

(b) (6 points) Build a complete decision tree by maximizing the information gain at each node split.

There are 3 possible splits:

- Experience ≤ 9 versus Experience > 9:
  Information gain $= 0.97095 - 0 - \frac{8}{10}(-\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8}) = 0.97095 - \frac{8}{10}(0.811278) = 0.97950 - 0.64902 = 0.32193$.
- Experience ≤ 13 versus Experience > 13:
  Information gain $= 0.97095 - \frac{4}{10}(1) - \frac{6}{10}(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}) = 0.97095 - \frac{4}{10} - \frac{6}{10}(0.918295) = 0.97950 - 0.95098 = 0.019973$.
- Experience ≤ 16.5 versus Experience > 16.5:
  Information gain $= 0.97095 - \frac{6}{10}(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}) - 0 = 0.97095 - \frac{6}{10}(0.918295) = 0.97950 - 0.55078 = 0.41998$.

The third split at 16.5 is the best. The decision tree is:

```
if  Experience > 16.5 then Success
else  if  Experience <= 9 then Failed
      else if Experience <= 13 then Success
           else Failed
```

(c) (2 points) Suppose another piece of information is available so that the samples might be easier classified. The second input attribute is "University" with four possible values:

- A if the programmer is a graduate of the NUS,
- B if the programmer is a graduate of the NTU,
- C if the programmer is a graduate of the SMU,
- D if the programmer is a graduate of a non-local university.

What is the number of possible splits that need to be considered using this attribute? What are these splits?

This new attribute is categorical, hence there are $2^{N-1} - 1 = 7$ splits:

i. A vs BCD

ii. B vs ACD

iii. C vs ABD

iv. D vs ABC

v. AB vs CD

vi. AC vs DB

vii. AD vs BC

3. (10 points) You are given the following six data points:

$$\mathbf{x_1} = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix},\ \mathbf{x_2} = \begin{pmatrix} 4 \\ -0.5 \\ 1 \end{pmatrix},\ \mathbf{x_3} = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$\mathbf{x_4} = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix},\ \mathbf{x_5} = \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix},\ \mathbf{x_6} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

A linear unit perceptron is trained to separate the data points: $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$ belong to class 0, while the remaining data points belong to class 1.

(a) (2 points) Given the weight vector $\mathbf{w_0} = \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$, what is the accuracy of the perceptron with the following classification rule:

`If output > 0, predict Class 1, else predict Class 0.`

- $o_1 = \mathbf{x_1}^T\mathbf{w_0} = -1.5$, predict Class 0
- $o_2 = \mathbf{x_2}^T\mathbf{w_0} = -2.75$, predict Class 0
- $o_3 = \mathbf{x_3}^T\mathbf{w_0} = -2.5$, predict Class 0
- $o_4 = \mathbf{x_4}^T\mathbf{w_0} = 0.5$, predict Class 1
- $o_5 = \mathbf{x_5}^T\mathbf{w_0} = -0.5$, predict Class 0
- $o_6 = \mathbf{x_6}^T\mathbf{w_0} = 0$, predict Class 0

Accuracy = $4/6 = 66.67\%$.

(b) (4 points) Starting from the initial point $\mathbf{w_0} = \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$, compute $\mathbf{w_1}$ using the gradient descent rule with learning parameter $\eta = 0.01$ and **batch** update to minimize the sum of squared errors:

$$\frac{1}{2}\sum_{i=1}^{6}(t_i - o_i)^2$$

where t_i is the class label and o_i is the output from the preceptron.

$$\begin{aligned}
\sum_{i=1}^{6}(t_i - o_i)\mathbf{x_i} &= (0-(-1.5))\begin{pmatrix}3\\1\\1\end{pmatrix} + (0-(-2.75))\begin{pmatrix}4\\-0.5\\1\end{pmatrix} + (0-(-2.5))\begin{pmatrix}3\\-1\\1\end{pmatrix} \\
&\quad +(1-(0.5))\begin{pmatrix}1\\3\\1\end{pmatrix} + (1-(-0.5))\begin{pmatrix}3\\3\\1\end{pmatrix} + (1-(0))\begin{pmatrix}2\\3\\1\end{pmatrix} \\
&= \begin{pmatrix}30\\6.625\\9.75\end{pmatrix} \\
\mathbf{w_1} &= \begin{pmatrix}-\frac{1}{2}\\\frac{1}{2}\\-\frac{1}{2}\end{pmatrix} + 0.01\begin{pmatrix}30\\6.625\\9.75\end{pmatrix} \\
&= \begin{pmatrix}-0.2\\0.56625\\-0.40250\end{pmatrix}
\end{aligned}$$

(c) (2 points) Compute the sum of squared errors before and after weight update.

- SSE before weight update: $\frac{1}{2}\sum_{i=1}^{6}(t_i - o_i)^2 = 9.78125$
- After update weights:
  - $o_1 = \mathbf{x_1}^T\mathbf{w_1} = -0.4363$
  - $o_2 = \mathbf{x_2}^T\mathbf{w_1} = -1.4856$
  - $o_3 = \mathbf{x_3}^T\mathbf{w_1} = -1.5688$
  - $o_4 = \mathbf{x_4}^T\mathbf{w_1} = 1.0962$
  - $o_5 = \mathbf{x_5}^T\mathbf{w_1} = 0.6962$
  - $o_6 = \mathbf{x_6}^T\mathbf{w_1} = 0.8962$
- New SSE $= \frac{1}{2}\sum_{i=1}^{6}(t_i - o_i)^2 = 2.48533$.

(d) (2 points) What is the accuracy of the perceptron after the weight update?

100%.

4. (10 points) Consider the following hypothetical bank data on customers' use of credit card facilities.

	Years	Salary	Used credit
1.	4	43	0
2.	10	65	1
3.	1	53	0
4.	3	95	0
5.	15	88	0
6.	9	112	1
7.	8	70	0
8.	12	120	0

where the data attributes are

- Years: the number of years that a customer has been with the bank.
- Salary: customer's salary (in thousands of dollars).
- Used credit = 1 if the customer has left an unpaid credit card balance at the end of at least one month in the prior year, Used credit = 0 if balance was paid off at the end of each month.

We are interested in building a classifier to predict "Used credit".

A network that consists of only one neuron is trained using the data set. Scale the values of "Years" by dividing them by 10 and the values of "Salary" by dividing them by 100 before answering the questions below. Starting from the initial weight

$$\mathbf{w_0} = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix},$$

compute $\mathbf{w_1}$ using the *incremental* gradient descent rule with the learning rate $\eta = 1$ and the network's output computed as the sigmoid of the weighted inputs.

Note: Consider only data sample #1.

- **Weighted output =** $w_1x_1 + w_2x_2 = 0.5 \times 0.4 - 0.5 \times 0.43 = -0.015$**.**
- **Network output =** $o_1 = \sigma(-0.015) = 1/(1 + e^{0.015}) = 0.4963$**.**
- **Error signal =** $\delta_1 = (t_1 - o_1) \times (o_1)(1 - o_1) = (0 - 0.4963)(0.4963)(0.5037) = -0.1240$**.**
- **Updated weight =** $\mathbf{w_1} = \mathbf{w_0} + 1 \times \delta_1 \times \mathbf{x^1} =$

$$\begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix} - 0.1240 \times \begin{pmatrix} 0.4 \\ 0.43 \end{pmatrix} = \begin{pmatrix} 0.4504 \\ -0.5533 \end{pmatrix}$$