COMPUTER ENGINEERING DEPARTMENT

SUBJECT: MACHINE LEARNING

COURSE: T.E. **YEAR: 2020-2021** **SEMESTER: VI**
DEPT: COMPUTER ENGINEERING
SUBJECT CODE: CSDLO6021 **EXAMINATION DATE: 11/06/2021**

MACHINE LEARNING
ANSWER SHEET

NAME	**:**	AMEY MAHENDRA THAKUR
SEAT NO.	**:**	61021145
EXAM	**:**	SEMESTER VI
SUBJECT	**:**	MACHINE LEARNING
DATE	**:**	11-06-2021
DAY	**:**	FRIDAY

STUDENT SIGNATURE: Amey

Q2 A]

i)

Support Vector Machine

- A Support Vector Machine (SVM) is a supervised learning algorithm that sorts data into two categories.
- A Support Vector Machine is also known as a Support Vector Network (SVN).
- It is trained with a series of data already classified into two categories, building the model as it is initially trained.
- An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.
- SVMs are used in text categorization, image classification, handwriting recognition and in the scienced.

Margin:

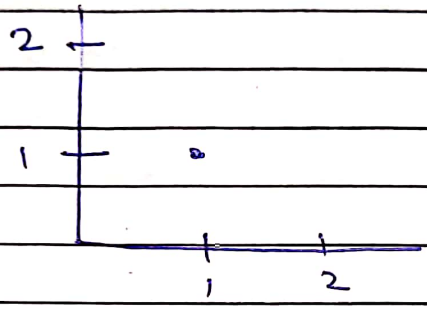
- A margin is separation of line to the closer class points.
- The margin is calculated as the perpendicular distance from the line to only the closest point.

Margin Boundary : $\frac{2}{\sqrt{w}}$

11 - 06 - 2021

STUDENT SIGNATURE: Amey

How to find margin (Example).



$y = x_1 + 2x_2 - 5.5.$

$a + 2a + b = -1$

$2a + 6a + b = 1$

$\therefore a = \frac{2}{5}, \quad b = \frac{11}{5}$

Optimal hyperplane is

$\overline{w} = (2/5, 4/5)$

and $b = -11/5$

Margin boundary is $2/|\overline{w}|$

$= 2/\sqrt{4/25 + 16/25} = 2/(2\sqrt{5}/5)$

$= \sqrt{5}$

Q2 A)

iii)

Steps for developing ML applications:

① Gathering data:

- This step is very important because the quality of data that you gather will directly determine how good your predictive model will be.
- We have to collect data from different sources for our ML application training purpose.
- This includes collecting samples by scraping a website and extracting data from an RSS feed or an API.

② Preparing the data:

- Data preparation is where we load our data into a suitable place and prepare it for use in our system for training.
- The benefit of having this standard format is that you can use mix and matching algorithms and data sources.

③ Choosing a model:

- There are many models that the data scientists and researcher have created over years.
- Some of them are well suited for image data, other for sequence and some for numerical data.
- It involves recognizing patterns, identifying outliers and detection of novelty

④ Training:

- In this step we will use our data to incrementally improve our models ability to predict the data we have inserted.
- Depending on the algorithm, feed the algorithm good clean data from previous steps and extract knowledge or information.
- The knowledge extracted is stored in a format that is readily usable by a machine for next steps.

⑤ Evaluation:

- Once the training is complete, it's time to check if the model is good for using evaluation.
- This is where testing datasets comes into play.
- Evaluation allows us to test our model against data that has never been used for training.

⑥ Parameter Tuning:

- Once we are done with evaluation, we want to see if we can further improve our training in any way.
- We can do this by tuning our parameters.

⑦ Prediction:

- It is a step where we get to answer for some questions.
- It is the point where the value of machine learning is realized

11 - 06 - 2021

STUDENT SIGNATURE: Amey

Q2 B

i)

Sol^n:

We will calculate split for all attributes i.e. Income, Defaulting, Creditscore, Location.

Income:

$$\text{Split} = \frac{5}{14}\,\text{gini}(\text{Low}) + \frac{4}{14}\,\text{gini}(\text{High}) + \frac{5}{14}\,\text{gini}(\text{Medium})$$

$$= \underline{0.392}$$

Defaulting:

$$\text{Split} = \frac{4}{14}\,\text{gin}(\text{High}) + \frac{6}{14}\,\text{gini}(\text{medium}) + \frac{4}{14}\,\text{gini}(\text{low})$$

$$= \underline{0.438}$$

Creditscore:

$$\text{Split} = \frac{7}{14}\,\text{gini}(\text{High}) + \frac{7}{14}\,\text{gini}(\text{Low}) = 0.493$$

$$= \underline{0.493}$$

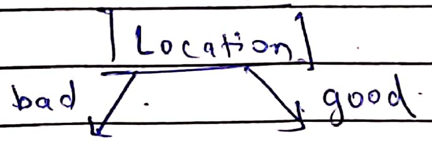
Location:

$$\text{Split} = \frac{8}{14}\,\text{gini}(\text{bad}) + \frac{6}{14}\,\text{gini}(\text{good})$$

$$= \underline{0.336}.$$

∵ Split value of location is smallest

∴ It will be root node



11-06-2021

STUDENT SIGNATURE: Amey

Now we will split bad branch considering remaining attributes

Income:

split $= \frac{3}{8}$ gini (low) $+ \frac{2}{8}$ gini (high) $+ \frac{3}{8}$ gini (medium)

$= 0.295$

Defaulting:

Split $= \frac{3}{8}$ gini (High) $+ \frac{3}{8}$ gini (medium) $+ \frac{2}{8}$ gini (low)

$= 0.37$

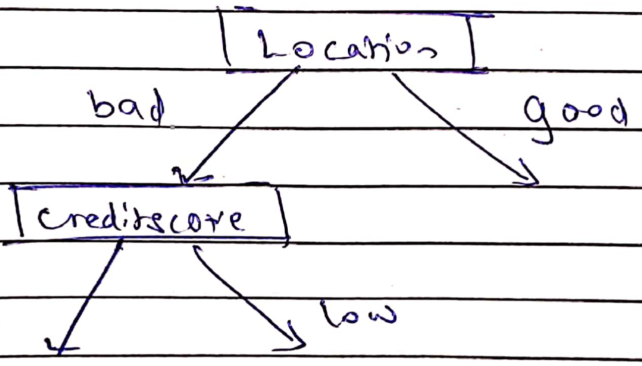
Creditscore:

Split $= \frac{4}{8}$ gini (High) $+ \frac{4}{8}$ gini (low).

$= 0.25$

∴ Split value of creditscore is smallest.

∴ Creditscore node is bad branch.



11 - 06 - 2021

STUDENT SIGNATURE: Amey

Now we will split good branch considering remaining attributes

Income:

$$\text{Split} = \frac{2}{6}\,\text{gini (low)} + \frac{2}{6}\,\text{gini (High)} + \frac{2}{6}\,\text{gini (medium)}$$

$$= 0.295$$

Defaulting:

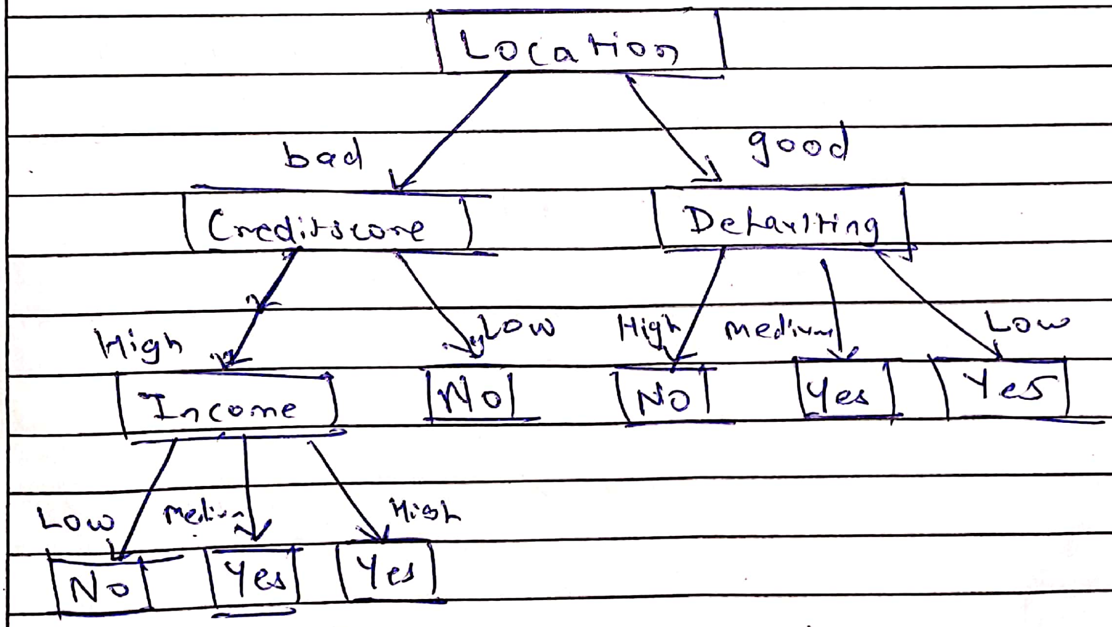
$$\text{Split} = \frac{1}{6}\,\text{gini (High)} + \frac{2}{6}\,(\text{medium}) + \frac{3}{6}\,(\text{Low})$$

$$= \underline{0}$$

∵ Split value of defaulting is smallest

∴ Defaulting will be node of good branch

∴ Decision Tree


Location
bad
good
Creditscore
Defaulting
High
Low
High
medium
Low
Income
No
No
Yes
Yes
Low
medium
High
No
Yes
Yes