CSIT375 AI for Cybersecurity

Assignment

Due Date: 20 August 2024

This assignment is for students taking CSIT375.

There are 3 questions in this assignment which contribute to 20% of the final mark. The questions must be solved manually, and the use of Python coding is not permitted.

Question 1 (8 marks)

- **Objectives**: In this question, we learn how to build an **unsupervised** anomaly detection model using **Gaussian distribution**. Please refer to the lecture notes for more details.
- **Tasks**: The dataset below presents information about the CPU usage and memory usage of a server over time. We need to identify any anomalous behaviours in the server's performance using the Gaussian Anomaly Detection technique.

Time (hours)	CPU Usage (%)	Memory Usage (%)
1	20	30
2	25	35
3	22	31
4	23	32
5	21	33
6	45	80
7	48	85
8	50	90
9	46	83
10	47	87

Complete the following subtasks:

1.1 (4 marks) Fit Gaussian distributions to the data by determining the mean and variance for each feature (CPU usage and memory usage). Please show all steps of your calculations.

1.2 (4 marks) The following data shows the server's performance at the 11th and 12th hours. Please identify whether these data points are normal or anomalous. Assume that the detection threshold is set to 0.0001.

Time (hours)	CPU Usage (%)	Memory Usage (%)
11	54	78
12	15	35

Question 2 (6 marks)

- **Objectives**: In this question, we learn how to build a **supervised** anomaly detection model using **Support Vector Machines** (SVMs). For more details, please refer to the example on the lecture notes.
- **Tasks**: Suppose we have a dataset consisting of the CPU usage and memory usage of a server over time. The corresponding labels are provided to indicate whether the server's behaviours are normal or anomalous. Note that, this dataset is not relevant to Question 1.

Time (hours)	CPU Usage (%)	Memory Usage (%)	Label
1	1	2	Normal
2	2	3	Normal
3	2	1	Normal
4	5	6	Anomaly
5	6	7	Anomaly
6	6	5	Anomaly

Complete the following subtasks:

2.1 (3 marks) Use the **geometric intuition** of SVMs to determine the support vectors.

2.2 (3 marks) Use the **geometric intuition** of SVMs to visually identify the optimal decision boundary and model parameters (i.e., $\boldsymbol{w}$ and b) without explicitly solving the optimisation problem. Draw the boundary and the data points to support your answer.

Question 3 (6 marks)

- **Objectives**: In this question, we learn how to build a malware threat detection model using **decision tree algorithms**. Please refer to the lecture notes for more details.
- **Tasks**: Suppose we are tasked with developing a malware threat detection system based on features extracted from files. The dataset below presents four **categorical** features of files, including file size, file type, number of sections, and suspicious keywords. Their corresponding labels are provided to indicate whether they are malware or not.

File Size	File Type	Number of Sections	Suspicious Keyword	Malware
100	Executable	3	5	Yes
100	Executable	1	2	No
100	Document	3	2	No
100	Document	1	5	Unknown
300	Executable	3	2	No
300	Executable	1	2	No
300	Executable	1	5	Unknown
300	Document	3	5	Yes
200	Executable	3	2	No
200	Executable	1	2	No
200	Document	3	5	Yes
200	Document	1	5	No

Use the **ID3 algorithm** to construct a decision tree from the training data and generate all rules for detecting malware. You must show all the results of computations. The following template can be used to show the computation of entropies.

	Entropy	Malware = Yes	Malware = Unknown	Malware = No
H(C| File Size=100) H(C| File Size=300) H(C| File Size=200)				
H(C| File Size)				
H(C| File Type=Executable) H(C| File Type=Document)				
H(C| File Type)				
H(C| Number of Sections=1) H(C| Number of Sections=3)				
H(C| Number of Sections)				
H(C| Suspicious Keyword=5) H(C| Suspicious Keyword=2)				
H(C| Suspicious Keyword)				

Submission Guidelines

- Your submission must be combined into a single PDF.
- Please name it as "*StudentID-CSIT375.pdf*".
- Please submit it via Moodle by the due date.

****END****