Assignment

Learning outcomes:

LO5: Explore potential distribution-shift sensitivity and the value of robust performance objectives.

LO6: Investigate formulations and algorithms that imbue robustness in an AI system.

Plagiarism declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. This assignment is my own work.

3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work.

4. I acknowledge that copying someone else's assignment (or part of it) is wrong and declare that my assignments are my own work.

Name:

1. Instructions and guidelines (Read carefully)

Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First Name Last Name Assignment – e.g., Lilly Smith Assignment. NB:** *Please ensure that you use the name that appears in your participant profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, "Start writing here" under each question. Please type your answer there.

3. Submit your assignment as a **Microsoft Word document only**. No other file types will be accepted.

4. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

PLEASE NOTE: Plagiarism cases will be investigated in line with the Terms and Conditions for Participants.

IMPORTANT NOTICE: Please ensure that you have checked your program calendar for the due date for this assignment.

Guidelines

1. There are 5 pages and 3 questions in this assignment.

2. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully but concisely and as directly as possible. Follow all specific instructions for individual questions (e.g., “list,” “in point form”).

3. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

2. Questions

Too often, machine learning models are developed in a laboratory setting without considering how well they are prepared for real-world deployment. Indeed, in the real world, many factors can undermine the robustness and reliability of ML models. In this assignment, consider how you would identify vulnerabilities in your ML model and mitigate the potential impact of these vulnerabilities.

> **Note:**
>
> As with the Module 1 assignment, consider how ML is or could be used in the context of your organization and discuss one (potential) application thereof. You may continue with the AI initiative used in your Module 1 assignment or choose an alternative ML project. Whichever you choose, please ensure that you provide sufficient context for the grader.

Question 1

Who could stand to gain from undermining your ML model? Briefly explain the model you have in mind – including the task it is intended to perform and how it should function once deployed in a real-world context – and discuss how potential adversaries might benefit from subverting it.

(Max. 250 words)

Start writing here:

Question 2

Simple oversights can arguably be as damaging as adversarial attacks. In this unit, you learned two key lessons about robustness: anticipating distribution-shift and accepting that there is a trade-off between performance and robustness. Explore the implications of these lessons in the context of your ML model by outlining how well the data and performance measure you intend to use represent reality, why it is necessary to be aware of limitations, and what you consider to be the non-negotiables (i.e., priors) for robust performance and their potential impact on accuracy.

(Max. 250 words)

Start writing here:

Question 3

Although robustness is never entirely guaranteed, the way a model is formulated and the chosen algorithms can imbue robustness. Discuss the steps your organization could take to minimize vulnerabilities and ensure acceptable performance for your ML model, and what you intend to achieve in the process.

(Max. 250 words)

Start writing here:

3. Rubric

	Criteria not met	Criteria met	Good	Exceptional
Question 1: *Adversarial threats*	No submission. OR The response does not explain an ML model or likely adversarial threats.	The response adequately explains the purpose of an ML model and potential adversarial threats.	The response thoroughly explains the purpose and real-world requirements for an ML model, as well as potential adversarial threats.	The response thoroughly explains the purpose and real-world requirements for an ML model and demonstrates a good understanding of the motivations and potential impact of likely adversarial threats.
Question 2: *Requirements for robust performance*	No submission. OR The response does not outline the training data, performance measure, or priors that will be used for the ML model.	The response adequately describes the training data, performance measure, and a prior for the ML model, but it lacks a clear understanding of acceptable trade-offs between accuracy and robustness.	The response analyzes the robustness of the training data and performance measure for the ML model and defines an acceptable prior.	The response thoroughly analyzes the robustness of the training data and performance measure for the ML model and clearly defines an acceptable prior.

Question 3: *Imbuing robustness*	No submission. OR The suggested process does not incorporate robust model development best practices.	The suggested process aligns with robust model development best practices, though the analysis of how these apply in the context of the ML model is lacking.	The response provides a clear process for imbuing robustness in the ML model, in line with robust development best practices.	The response provides a clear and comprehensive process for imbuing robustness in the ML model, including example formulations and algorithms in line with robust development best practices.