Mock Exam

Security and Privacy of Machine Learning (NWI-IMC069)

1. **What aspects of the CIA triad are compromised from the attacks on AI (evasion attack, poisoning attack, sponge poisoning attack, backdoor attack, model stealing attack, membership inference attack, and model inversion attack). As a reminder the CIA triad stands for:**

   - ***Confidentiality*: Only authorized parties can access information or services.**
   - ***Integrity*: Information is protected from unauthorized alteration (i.e., creation, modification, and deletion).**
   - ***Availability*: Information or services must be available to all authorized parties whenever they are needed.**

   A short explanation of each choice is required.

2. **What threat model shall we assume for the adversary when designing an attack, black box or white box? Why? Similarly, what threat model shall we use for an attacker when designing a defense? Why?**

3. Threat modeling:

   ◯ Uses models to identify security flaws.

   ◯ Uses implementation details.

   ◯ Provides theoretical guarantees for its outcomes.

   ◯ All of the above.

   ◯ None of the above.

4. **Explain the concept of federated learning and discuss its potential security vulnerabilities. What are the key challenges in ensuring the security and privacy of data in federated learning systems?**

5. **Describe a full threat model for an attack of your choice.**

6. **What are the main differences in the training of SL and FL (briefly explain)? What is the main advantage of using SL over HFL? what is the advantage of using Boomerang SL over Vanilla SL?**

7. Can we make an evasion attack in federated learning (FL)?

   ◯ No. Evasion attacks can only be made in centralized ML.

   ◯ Yes. Also, it is different from centralized ML.

   ◯ Yes. However, the setup is the same as in ML.

   ◯ No. In FL, only poisoning attacks are possible.

8. **Why do adversarial examples exist?**

9. **Describe the FGSM algorithm.**
10. **List 3 defenses against evasion attack and describe one of those.**
11. **What is the difference between MIA and AIA privacy attacks?**
12. **Describe how you need to rewrite the FGSM code below to change it into a PGD attack.**
    - Explain what parameters you would include/remove
    - List which lines of code you would modify
    - Explain what modifications you would perform
    - We do not expect exact code, but make a list with pseudocode and explanations of the code

```python
class FGSM(Attack):

    def __init__(self, model, eps=8/255):
        super().__init__("FGSM", model)
        self.eps = eps
        self.supported_mode = ['default', 'targeted']

    def forward(self, images, labels):
        images = images.clone().detach().to(self.device)
        labels = labels.clone().detach().to(self.device)

        if self.targeted:
            target_labels = self.get_target_label(images, labels)

        loss = nn.CrossEntropyLoss()

        images.requires_grad = True
        outputs = self.get_logits(images)

        # Calculate loss
        if self.targeted:
            cost = -loss(outputs, target_labels)
        else:
            cost = loss(outputs, labels)

        # Update adversarial images
        grad = torch.autograd.grad(cost, images,
                                   retain_graph=False, create_graph=False)[0]

        adv_images = images + self.eps*grad.sign()
        adv_images = torch.clamp(adv_images, min=0, max=1).detach()
```

```
32
33              return adv_images
```