人工智能安全第二次作业

- 姓名：
- 学号：JY

思考题1

题干：干净标签投毒和脏标签投毒各有什么优缺点？前者一定比后者好吗？

回答如下：

干净标签投毒

干净标签投毒是运用在原始（像素）构成上和标签A的数据相近，但是在某层特征空间中和标签B的的特征相近的投毒数据，从而导致模型错误判断的攻击方法。具有以下特点：

- 中毒图像标签与视觉感官一致
- 通过少量投毒，攻击者可以使选定的目标测试图像被错误分类
- 受害模型仍然保有较高的分类准确率

优点包括：

- 可塑性高：攻击者通过调整投毒数据的噪声强度、范围和方向等参数，从而使得攻击具有很强的可塑性，攻击的变化范围很大。
- 检测难度大：因为干净标签投毒攻击与原数据相比，没有没有明显噪声或者失真，攻击数据与原数据相似程度很高，难以在数据清洗的时候被发现。
- 攻击面广：可以用于任何类型的机器学习模型和任务，包括计算机视觉和自然语言处理等多个领域。这使得攻击者可以在不同场景下使用干净标签投毒攻击来实施攻击。

缺点包括：

- 攻击样本制造困难：干净标签投毒攻击需要攻击者非常要求了解各种机器学习模型，在选择特征空间、评估攻击效果等方面能够实现参数的精准控制。攻击样本的制造过程比随机翻转这种方法复杂很多。
- 样本数据需求量大：干净标签投毒攻击需要攻击者拥有大量的带标记（即有良好的标签）的训练数据，才能制造出有效的攻击样本。如果攻击者没有足够的标记数据，那么攻击的效果会很差。

脏标签投毒

脏标签投毒通过注入大量的脏标签数据来扰乱模型的训练过程，使模型产生严重偏差和误判，从而达到干扰模型、破坏模型稳定性的目的。

优点包括：

- 脏标签制造方法灵活：有诸如“基于标签反转”、“基于优化的数据投毒”、“基于梯度的数据投毒”等许多方法制造脏标签
- 对攻击模型有灵活性：既有白盒攻击，通过梯度下降法训练制造较好的投毒数据；也可以做如标签反转的黑盒攻击
- 攻击成本具有低廉型：可以通过随机生成噪声进行脏标签攻击，攻击成本低

缺点包括：

- 伪装性差：攻击注入的标签数据与真实数据分布差异很大，且其分布规律通常很难与真实数据相似，容易被数据清洗

- 攻击效果不稳定：脏标签攻击是通过注入带有误导性的标签数据来干扰模型，其攻击效果往往是受数据分布、标签噪声等因素影响，攻击效果不稳定
- 误伤率高：脏标签攻击阈值较低，对整个训练集造成影响大，影响模型的整体性能，从而被发现的概率较高

前者不一定比后者好。

- 对目标模型不了解，无法针对性进行干净标签攻击的时候可以考虑进行大规模脏标签攻击
- 脏标签攻击能够以较为低廉的成本来完成攻击，除非模型数据清洗能力很强。

思考题2

题干：基于K-NN的中毒数据检测有什么优缺点？这种检测方法对于干净标签数据投毒和脏标签数据投毒攻击都有效吗？

基于K-NN的中毒数据检测是一种基于邻近度的异常检测方法，其主要思想是通过计算每个样本点与其最近的K个邻居之间的距离来判断该样本点是否为异常点。利用k-NN算法，为训练数据集中的每个训练数据计算标签，如果计算出的标签与该数据的真实标签不一致，则认为这个训练数据被污染，将其从训练数据中移除。

对于干净标签数据投毒攻击，该方法能够有效地检测出其中的异常点，并准确地识别出中毒数据。在Peri, N., Gupta, N., Huang, W.R., Fowl, L., Zhu, C., Feizi, S., Goldstein, T., & Dickerson, J.P. (2019). Deep k-NN Defense Against Clean-Label Data Poisoning Attacks. ECCV Workshops.里面，当 k 值选择恰当时，该方法可有效去除训练集中 99% 以上的中毒数据，对干净标签中毒攻击的防御成功率高达 100%，且不会将大量良性数据误检为中毒数据。

对于脏标签的数据投毒攻击，该方法检测结果可能不够准确。原因：在脏标签数据投毒攻击中，攻击者注入的标签数据之间有很大的类别差异，从而导致样本点之间距离计算的不准确性，k-NN分类的准确性下降，影响了中毒数据的检测。

思考题3

题干：为了防范数据投毒攻击，你还能想到什么样的方法来提前预防这类恶意攻击对模型的可用性和完整性产生破坏？

基于训练数据检测的防御：

- **异常检测**：
  - **原理**：使用异常检测技术识别和移除训练数据中的异常样本。这些异常样本可能是由于错误标记、噪声数据或恶意注入而产生的，比如基于孤立森林的异常检测方法。
- **数据重采样**：
  - **原理**：重新采样数据集，剔除不一致或不可信的数据，或者通过过采样/欠采样等技术平衡数据集的类别分布，从而提高模型的稳健性和泛化能力。
- **去噪自编码器的方法**：
  - **原理**：通过编码器将输入数据映射到低维的表示（编码），然后再通过解码器将该低维表示重构回原始数据。在训练过程中，去噪自编码器会被设计成在输入数据中引入噪声，然后尝试学习去除这些噪声并恢复原始数据。

基于数据增强的防御：

- **数据扩充**：
  - **原理**：通过对原始训练数据进行随机变换或扰动，生成新的训练样本，以增加数据的多样性。选择适合任务和数据类型的数据增强技术，例如图像数据可以采用旋转、缩放、裁剪、平移、翻转、加噪声等操作；文本数据可以进行词语替换、插入、删除等处理。

- **目的**：使模型在训练过程中接触到更多种类和变化的数据，增强其泛化能力，降低过拟合风险，并提高对抗攻击的鲁棒性。

基于鲁棒训练的防御：

- **对抗训练**：
  - **原理**：在训练过程中，向模型输入对抗性样本，同时让模型优化以最小化对抗样本的损失。
  - **具体步骤**：在每次训练迭代中，生成对抗性样本，将其与原始样本一起输入模型，然后更新模型参数以减小对抗样本的损失函数。
  - **例子**：在图像分类任务中，使用对抗生成网络（GAN）生成对抗性样本，然后将这些对抗性样本与原始图像一起输入到分类模型中进行训练。模型在训练过程中逐渐学习到对抗样本的特征，并逐渐提高对抗攻击的鲁棒性。
- **防御性数据增强**：
  - **原理**：通过改变训练数据，向模型提供更多多样化的输入，包括对抗性样本。
  - **具体步骤**：对训练数据进行随机扰动或添加，生成对抗性样本，并将其用于训练模型。
  - **例子**：在文本分类任务中，对原始文本进行词语替换、插入或删除，生成对抗性文本样本，然后将这些对抗性样本与原始文本一起用于模型训练。模型通过学习包含对抗性文本的更广泛的训练数据，能够更好地捕捉文本中的关键信息，从而提高对抗攻击的抵抗能力。
- **模型集成**：
  - **原理**：结合多个模型的预测结果，通过投票或加权平均来减少单一模型的过度自信，从而提高对抗攻击的鲁棒性。
  - **具体步骤**：训练多个不同结构或初始化的模型，然后结合它们的预测结果作为最终输出。
  - **例子**：收集了包含各种动物图像的数据集，包括猫、狗、鸟等，分别使用CNN、SVM和决策树这三种不同类型的模型对数据集进行训练。每个模型学习图像特征并尝试区分不同的动物类别。训练完成后，对测试集中的每张图像进行预测。对于每张图像，每个模型都会给出一个预测结果。最后，采用投票或加权平均的方式，综合利用三个模型的预测结果来得出最终的分类决策。