**Introduction of Machine Learning Security**

**Seminar Report**

顾凯铭 (206164050055)

**Contents of Seminar?**

Firstly , Fabio Roli introduced SAlfer Lab for us , which is focused on ML , Pattern Recognization ,and the Application of data science . Two main topics of the talk is adversarial ML lessons and future challenge. From the side of AI , AI has better performances than human in images recognization , by the use of imageNet. The relationship between high accuracy and rubusteness has been talken , if the dataset has been modified very slightly , high accuracy may not lead to high rubusteness. He also considered the smoothness assumption : points close to each other are more likely to share a label and how do these adversarial sample work. The professor taught the Evasion of nonlinear classifiers by give handwritten digits changing from 3 to 7 as a example. Beside the evasion attacks , we can form 6 different attack formats by considering attacker’s capability and attacker’s goal , if we only know feature representation and algorithm possibly , we can use surrogate data to classifier and get the feedback labels. The attack is more hard when attacks are constrained and invariant. The two main topics of future challenges is why machine learning is so vulnerable and regression of machine learning. The reasons of vulnerable ml is the large gradient and spurious correlation . And the regression of machine learning means even if the model has been re-trained , we can find new vulnerable regions which don’t appear in the previous model.

**What did I learn?**

From this seminar, i get in touch with the regression of machine learning firstly and know that retraining is a must for modern machine learning. Thanks to the slides of professor , i get that the current model is vulnerable to adversarial examples that the previous model was able to detect , which is called security regression. AI evaluation is not enough , and we should make an effort to grub the true meaning of machine learning model. Furthermore, I realized that evaluating AI solely through standard metrics is insufficient. We need to delve deeper to truly grasp the underlying mechanics and implications of machine learning models. This involves not just assessing performance through typical evaluation criteria but also understanding the broader context, potential vulnerabilities, and ensuring robust and secure model behavior.