Inferring Class Label Distribution of Training Data from Classifiers: An Accuracy-Augmented Meta-Classifier Attack

Paper

Raksha Ramakrishna, György Dán, KTH Royal Institute of Technology, Stockholm, Sweden

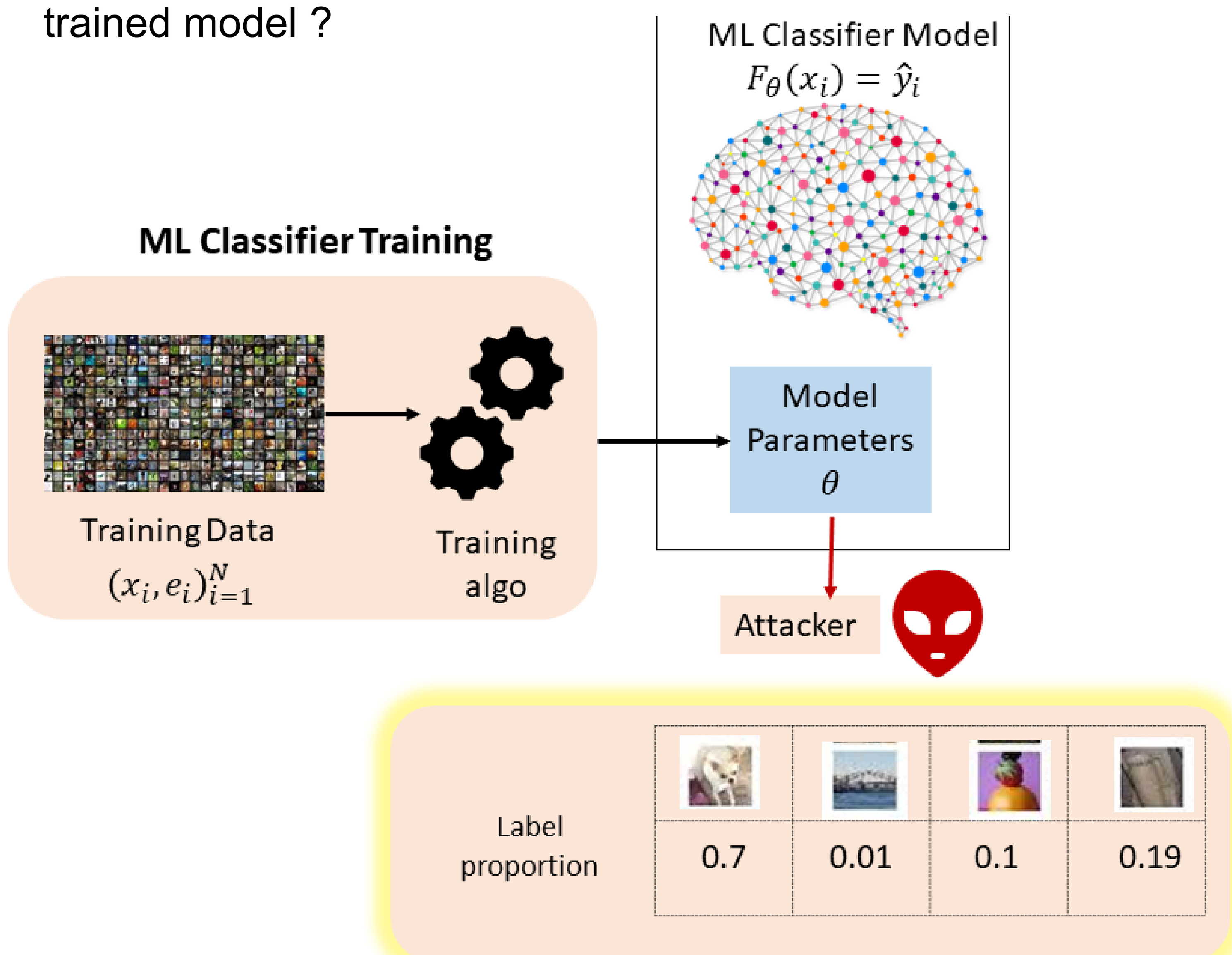
{rakshar, gyuri}@kth.se

Class-Label Distribution Inference

Class-label distribution of training data:

$$p = \frac{N_c}{\sum_{c=1}^{C} N_c}$$

- Sensitive information in banking, manufacturing, health
- Question: Can adversary infer class label distribution based on trained model ?



Attack Model:

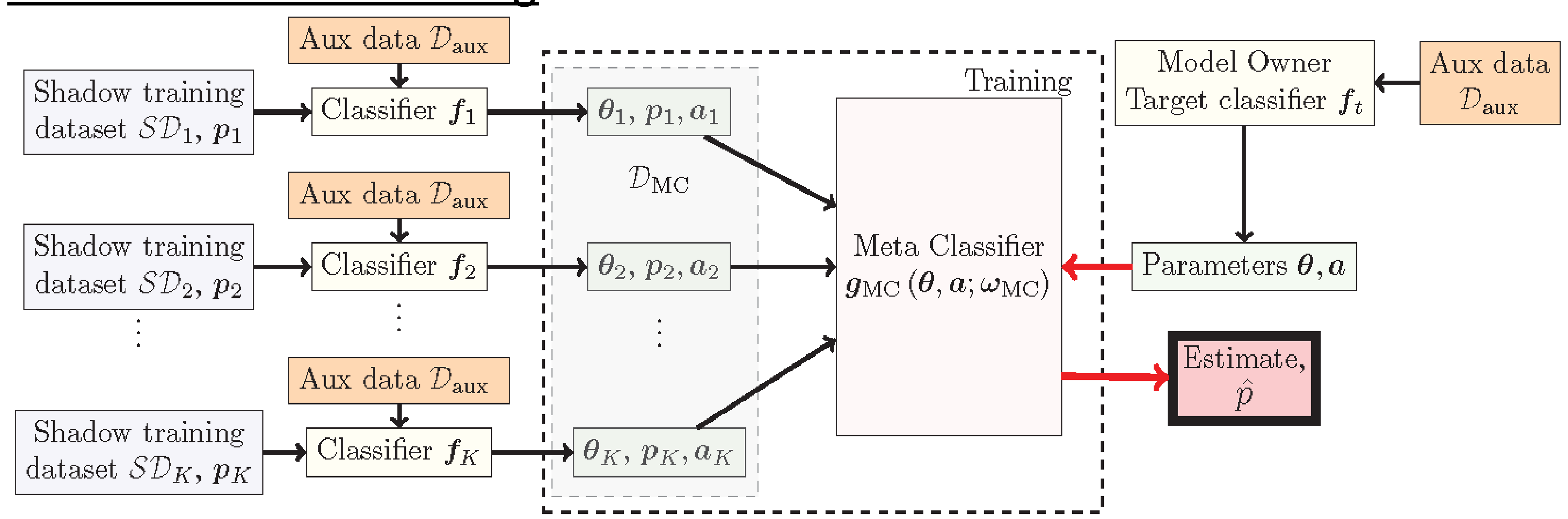
- White box attack
- Similar datasets available to attacker

Ingredients for class-label distribution inference:

- Target classifier $y = f_t(x; \theta)$ trained on class-label distribution p_t
- For inference: parameter θ, accuracy a over auxiliary dataset D_{aux} with N_{aux} samples from each class used
- Metric of accuracy

$$D_{KL}(p_t || \hat{p}) = \sum_{c=1}^{C} [p_t]_c \log\left(\frac{[p_t]_c}{[\hat{p}_t]_c}\right)$$

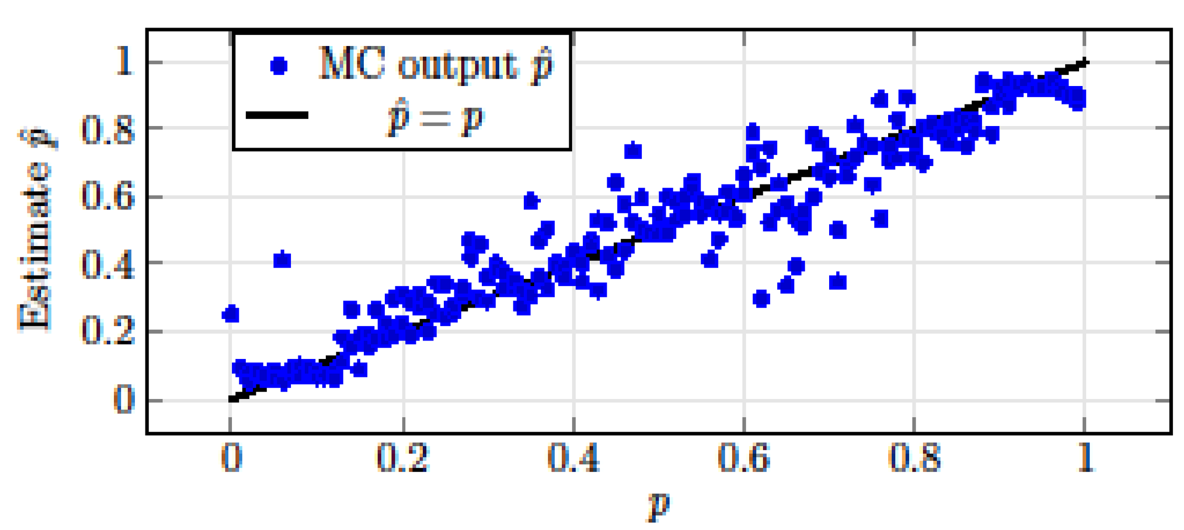
Meta-Classifier Training

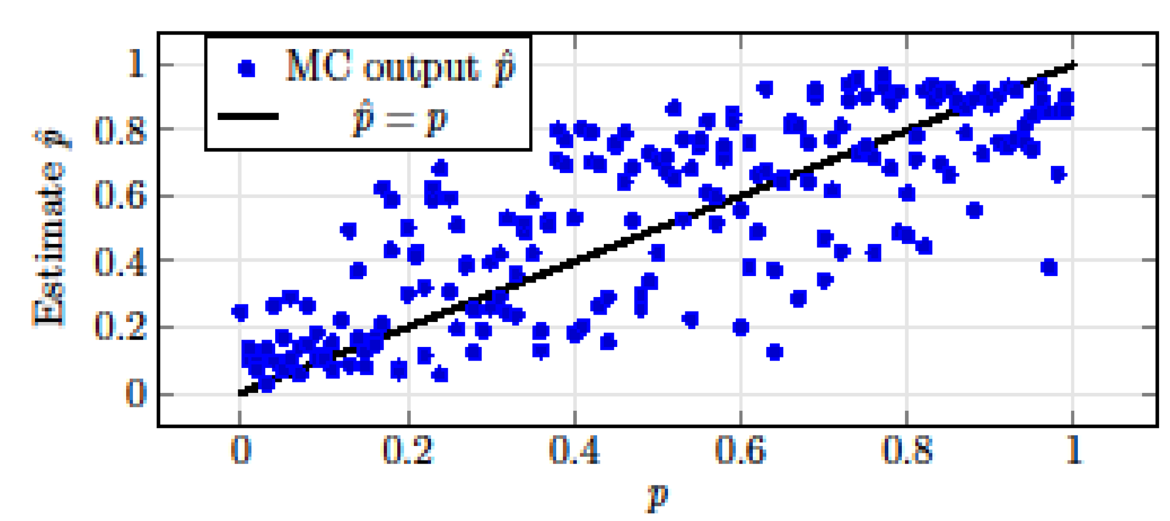


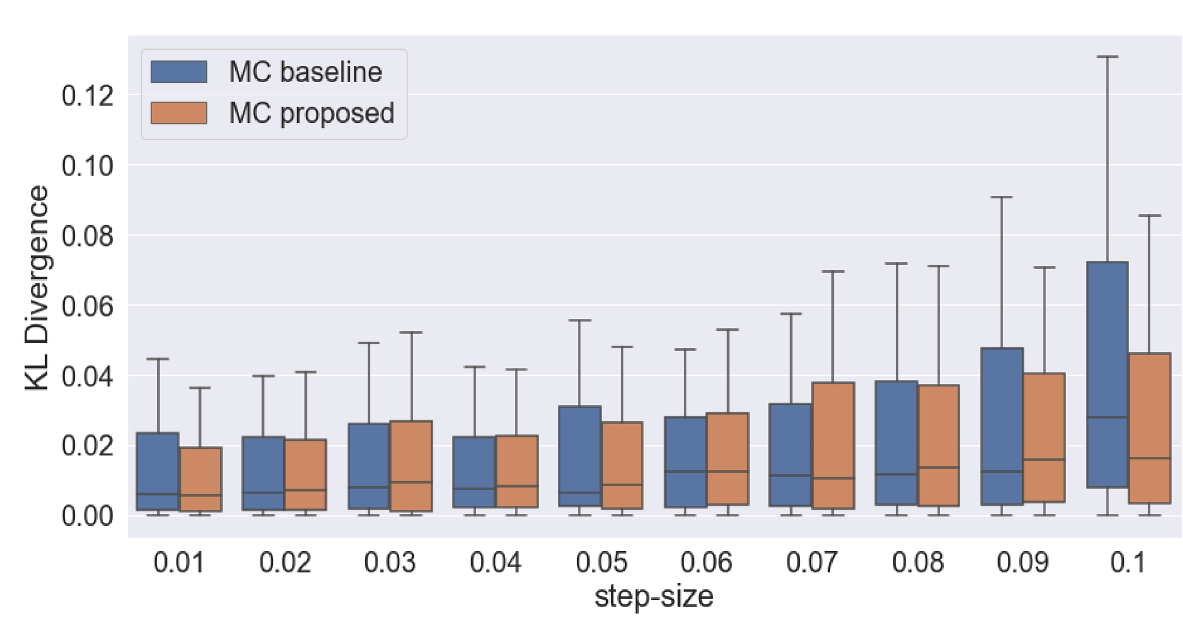
- Meta classifier $\hat{p} = g_{MC}(\theta, a; \omega_{MC})$ parameters learned via shadow-training
- Shadow-training datasets $\{SD_k, p_k\}$ to train shadow classifiers f_k
- Use parameters $\{\theta_k\}$ and accuracy $\{a_k\}$ to learn meta-classifier parameters ω_{MC}

Numerical Results

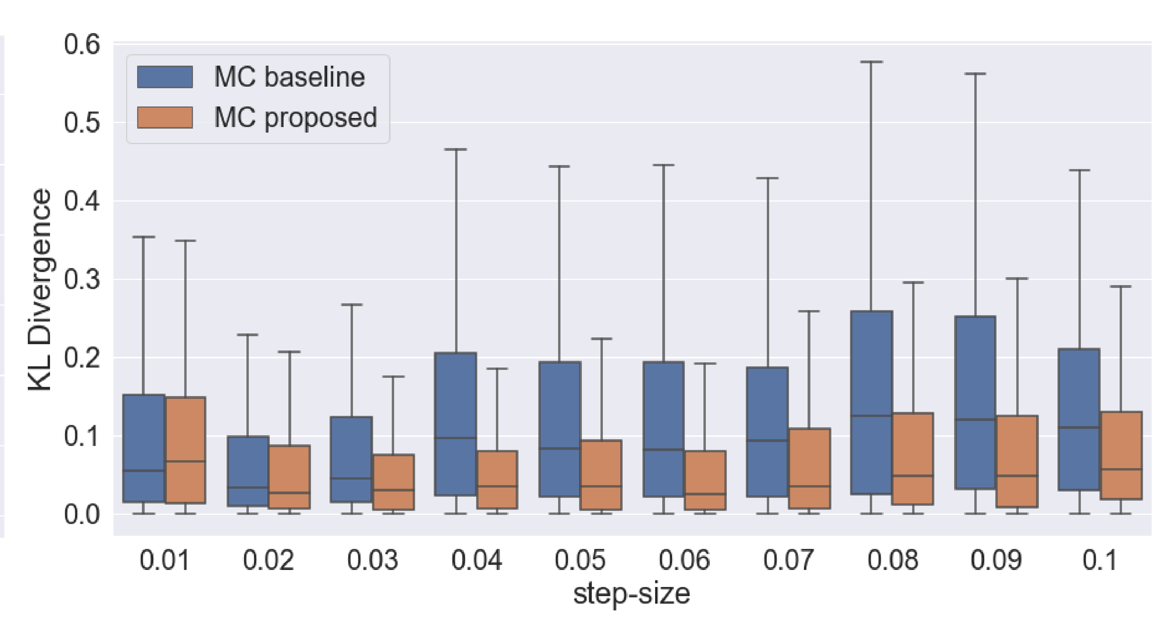
- Binary classification: Class-label distribution is Bernoulli (p)
- Accurate estimates for most values of p especially very imbalanced datasets
- Shadow-classifiers trained using class-label distributions with different step sizes (Δp)
- Large improvements over baseline (Ganju et.al)







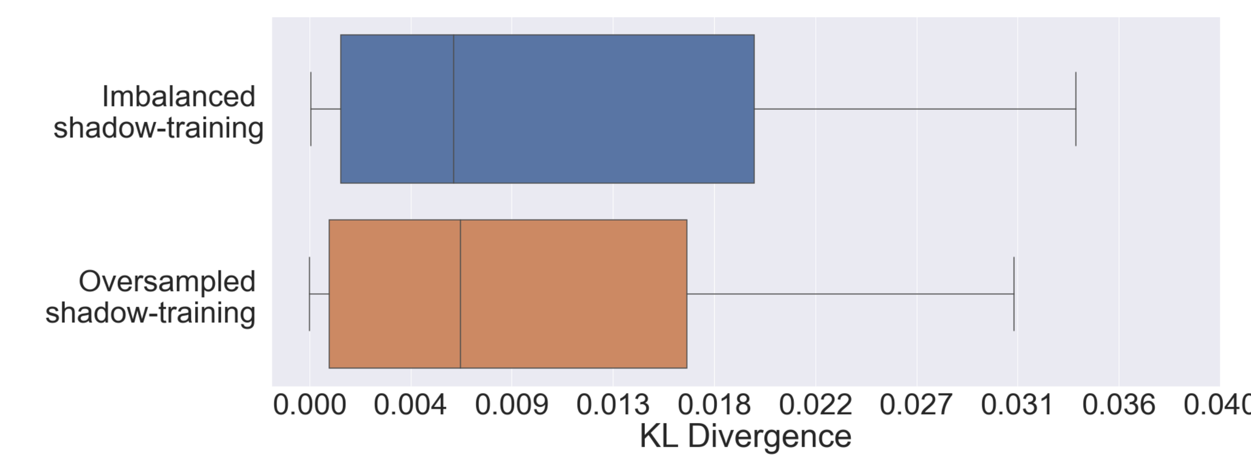
UCI Census Income Classification

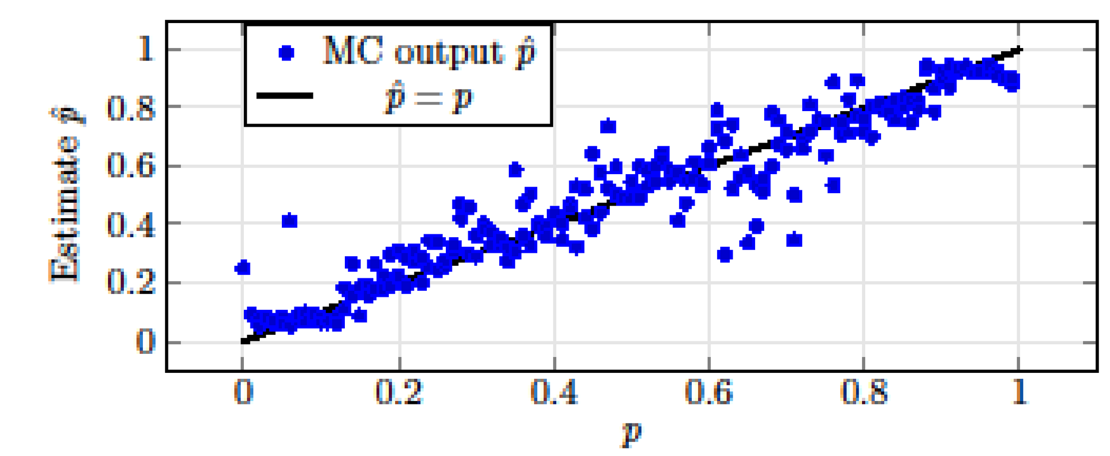


MNIST (0 and 1)

Potential Countermeasure

- Random oversampling of minority class: address class-imbalance
- Makes class-label distribution uniform
- Meta-classifier can still estimate original distribution!
- Further training on oversampled datasets improves performance





Future Work

- Develop new meta-classifiers specific to other target architectures like CNNs
- Mitigation measures
- Extension to Federated Learning

References

- G. Ateniese et al., "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," International Journal of Security and Networks, vol. 10, no. 3,2015.
- Ganju et al., "Property inference attacks on fully connected neural networks using permutation invariant representations", in Proc. of ACM Conf. on Computer and Communications Security (CCS), 2018.
- Rigaki et al., "A survey of privacy attacks in machine learning," arXiv preprint arXiv:2007.07646, 2020.
- Raksha Ramakrishna and György Dán., "Inferring Class-Label Distribution in Federated Learning. In Proc of 15th ACM Workshop on Artificial Intelligence and Security (AISec'22). ACM,pp. 45–56.