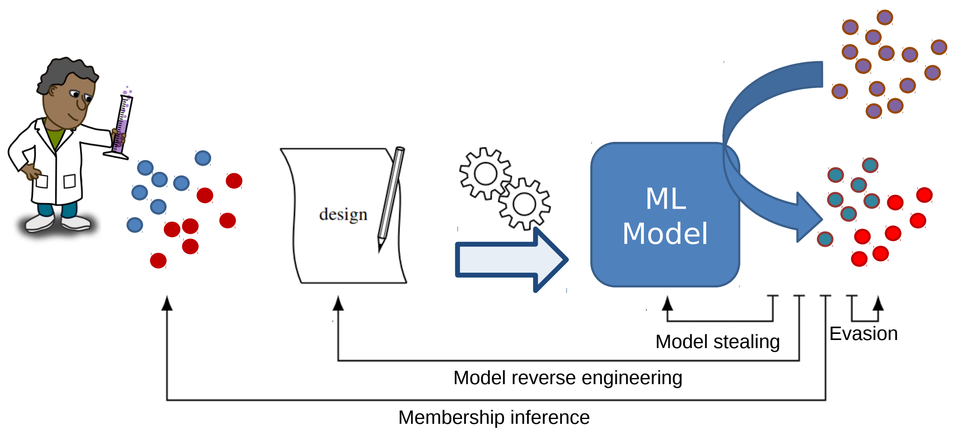
Killing four Birds with one Gaussian Process: The relation between different Test-time attacks

Kathrin Grosse, Michael T. Smith, Michael Backes

CISPA Helmholtz Center for Information Security / Saarland Informatics Campus, Sheffield University, CISPA Helmholtz Center for Information Security
Kathrin.grosse@cispa.saarland

Adversarial ML (test time attacks)



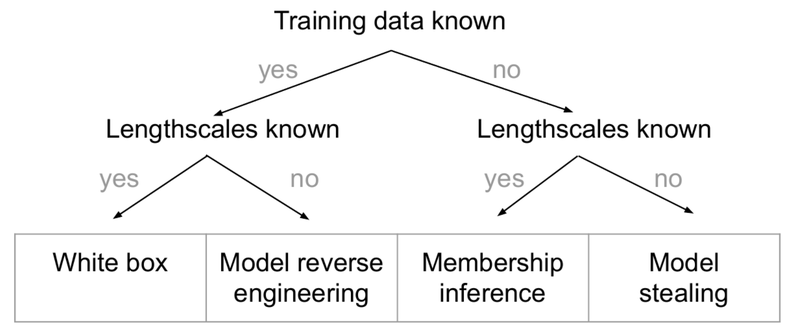
Why Gaussian processes?

GP, after training, **are fully specified** and **deterministic**

GP's **curvature** can be set using the **lengthscale**

GP are applied in **medical** settings, risk assessment is **crucial**

GP allow to **relate** IP based attacks:



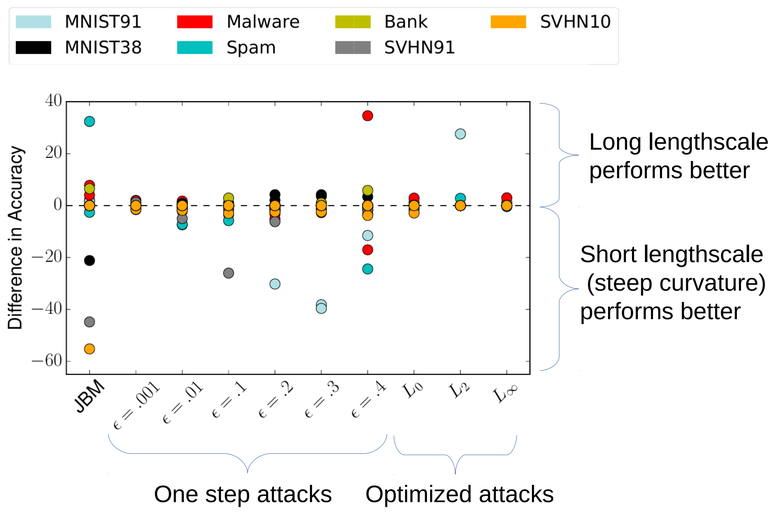
Evasion

Test **transferred** adversarial examples from **DNN, SVM, GP**

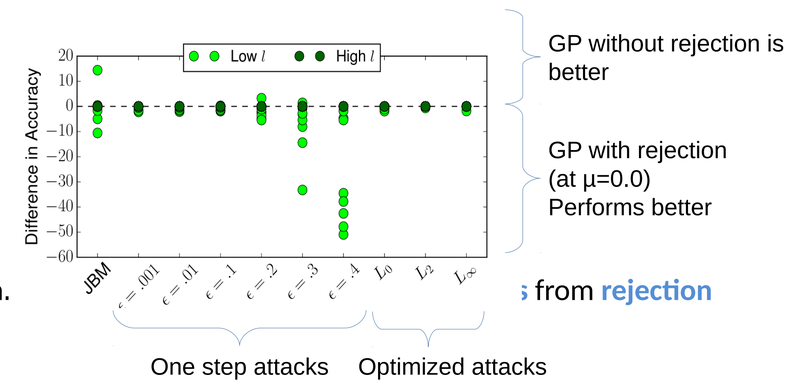
Compare long and short lengthscales (**steep** and **low curvature**)

Steep curvature is **harder** to attack with **one step attacks**

Low curvature is **harder** to attack with optimized **attacks**



Compare long and short lengthscales (**steep** and **low curvature**) and **reject** data if **output of GP is 0**



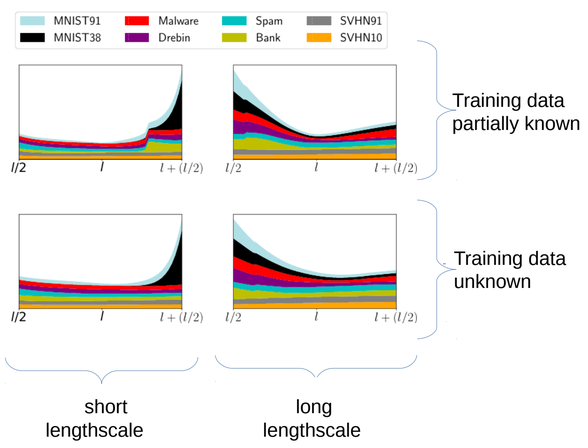
On. s from **rejection**

Model reverse engineering - lengthscale

Compare long and short lengthscales (**steep** and **low curvature**)

Try to infer **lengthscale** given **partial or no access** to used training data

A **short lengthscale** conceals the lengthscale **better**



Model reverse engineering - kernel

Compare long and short lengthscales (**steep** and **low curvature**)

Attempt to infer **kernel** used in GP

Attack is successful **regardless** of **curvature in RBF kernel**

	MNIST91	MNIST38	Malware	Drebin	Spam	Bank	SVHN91	SVHN10
RBF_S	✓	✓	✓	✓	✓	✓	✓	✓
RBF_L	✓	✓	✓	✓	✓	✓	✓	✓
Linear	×	×	×	/	✓	✓	✓	✓
Poly	✓	✓	✓	/	×	✓	✓	✓

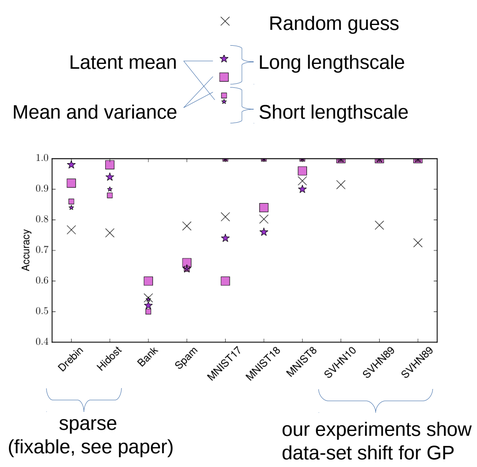
✓ Attack succeeds
× Attack fails
/ GP does not converge for kernel

Membership inference

Compare long and short lengthscales (**steep** and **low curvature**)

Try to infer if point is **in training data** given **latent mean / mean and variance**

A long lengthscale **is more robust** towards **membership inference**



Conclusion

AML attacks **should not** be studied **in isolation**.

Defending one attack might **increase** vulnerability for an **unrelated** attack!

- A **short** lengthscale is harder to attack with optimized attacks
- A **short** lengthscale conceals the lengthscale better
- Attack is successful **regardless of curvature** in RBF kernel
- A **long** lengthscale is more robust towards membership inference

Sources: Model reverse engineering: Oh et al. "Towards reverse-engineering black-box neural networks." Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 2019. 121-144.
Membership Inference: Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.
Evasion: Laskov, Pavel. "Practical evasion of a learning-based classifier: A case study." 2014 IEEE symposium on security and privacy. IEEE, 2014.

The University Of Sheffield.

UNIVERSITÄT DES SAARLANDES

CISPA HELMHOLTZ-ZENTRUM i. G.