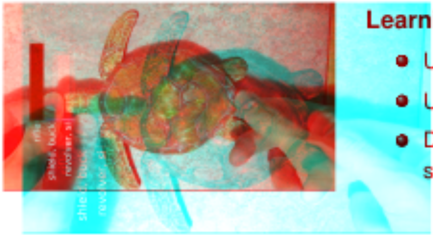
Interpretable Machine Learning

Local Explanations: Adversarial Examples



Learning goals

- Understand the definition of ADEs
- Understand first methods that generate ADEs
- Discuss potential causes of ADEs and standard defenses against them

ADVERSARIAL MACHINE LEARNING

- What happens if a computer system gets an erroneous input?
- Even worse:
  What happens if someone feeds in a malicious input on purpose to attack a system?

⇝ **Robustness** is important to ensure a safe service!

- **Adversarial ML** studies the robustness of machine learning (ML) algorithms to malicious input
- Two different kinds of attacks:
  - **Evasion attacks** mislead an employed ML model with manipulated inputs (our focus)
  - **Data Poisoning**: Malicious inputs to the training dataset

ADVERSARIAL MACHINE LEARNING

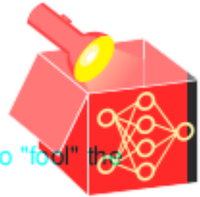
- What happens if a computer system gets an erroneous input?
- Even worse:
  What happens if someone feeds in a malicious input on purpose to attack a system?

⇝ **Robustness** is important to ensure a safe service!

- **Adversarial ML** studies the robustness of machine learning (ML) algorithms to malicious input
- Two different kinds of attacks:
  - **Evasion attacks** mislead an employed ML model with manipulated inputs (our focus)
  - **Data Poisoning**: Malicious inputs to the training dataset

Interpretable Machine Learning – 1 / 14

ADVERSARIAL EXAMPLES

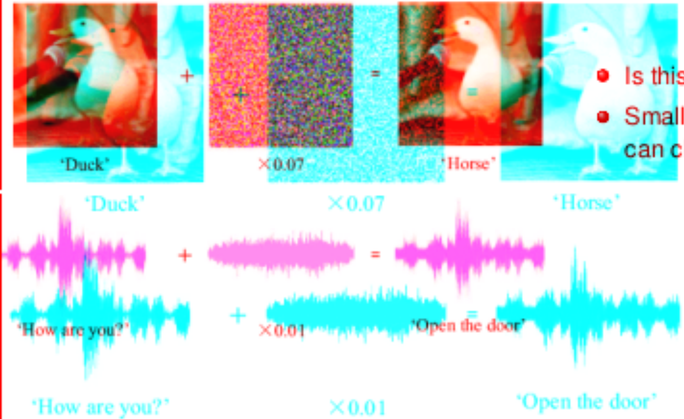


- **Informal Definition:** An ADE is an input to a model that is deliberately designed to "fool" the model into misclassifying it
- Even possible with low generalization error
- Both deep learning models (e.g., CNNs) and classical ML can be vulnerable to such attacks
- ADEs created from a real data observation **x** can be indistinguishable from **x** by a human observer
- Since the model misclassifies this input, it does not seem to have a real understanding of the underlying concepts of the provided inputs

Interpretable Machine Learning – 2 / 14

EXAMPLES: MODEL-ATTACKS

Gong & Poellabauer 2018



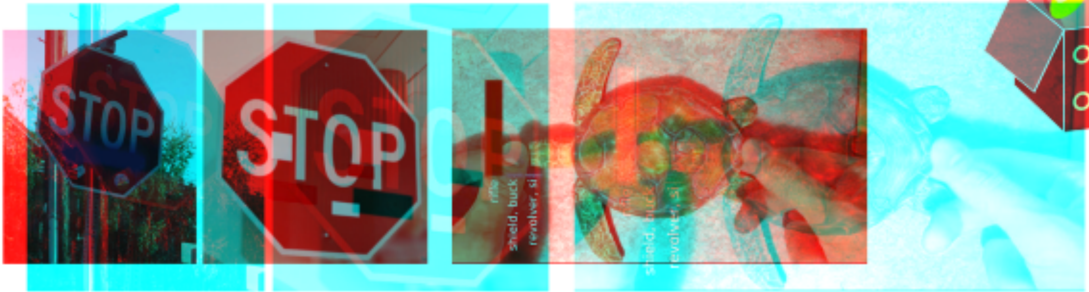
- Is this a duck or a horse?
- Small (hard-to-see) noise can change the prediction

- Is this a duck or a horse?
- Small (hard-to-see) noise can change the prediction

Interpretable Machine Learning – 3 / 14

EXAMPLES: IMAGE DATA

▸ Eykholt et al. (2018) ▸ Athalye et al. (2018)



- Stop signs can be missclassified
  e.g., because of graffiti
- With some well-placed patches, the model identifies it as a "right of way" sign
- 3D-print of a turtle
- Misclassified as a rifle (from every angle)
- Video: ▸ MITCSAIL (2017)

- Stop signs can be missclassified
  e.g., because of graffiti
- With some well-placed patches, the model identifies it as a "right of way" sign
- 3D-print of a turtle
- Misclassified as a rifle (from every angle)
- Video: ▸ MITCSAIL (2017)

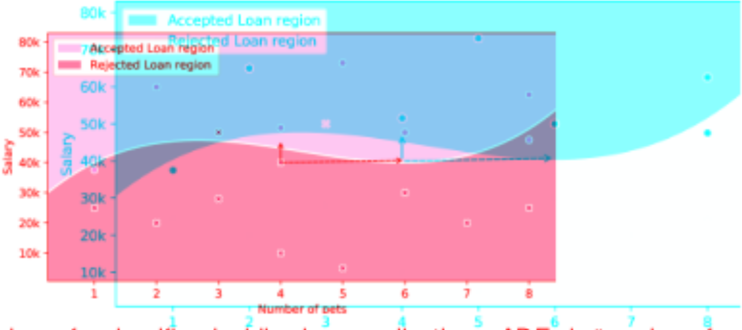
Interpretable Machine Learning – 4 / 14

EXAMPLE: TABULAR DATA

Ballet (2019)

What is imperceptibility on tabular data?

- Idea: experts focus on the most important features in their judgment
- An ADE arises from manipulating features the model deems important but experts do not



Decision boundary of a classifier deciding loan applications. ADE via “number of pets”

ADE AND INTERPRETABILITY

1. ADEs show where models fail ⇝ improved model understanding
2. Because of ADEs, we need more interpretability
3. Interpretation can lead to robustness against ADEs
4. Explanations can be used to construct ADEs (e.g., see numer of pets on previous slide)

Interpretable Machine Learning – 6 / 14

FORMAL DEFINITION

Adversarial Input

Let $\epsilon > 0$, $f : \mathcal{X} \rightarrow \mathcal{Y}$ be an ML model and $\mathbf{x} \in \mathcal{X}$ be a real data point that is correctly classified: $f(\mathbf{x}) = y_{\mathbf{x},true}$.

We call $\mathbf{a_x}$ an **adversarial input** to $\mathbf{x}$ if:

$$\|\mathbf{a_x} - \mathbf{x}\| < \epsilon \text{ and } f(\mathbf{a_x}) \neq y_{\mathbf{a_x},true} = f(\mathbf{x})$$

- $\mathbf{a_x}$ an is a data point close to a real, correctly classified input that is misclassified
- $\mathbf{a_x}$ is called **targeted** if the class it is assigned to is determined $f(\mathbf{a_x}) = y'$ with y' being a desired prediction
- Can be generalized to regression problems

Interpretable Machine Learning – 7 / 14

WHY DO ADE EXIST?

Non-exhaustive list of hypotheses:

1. **Low-probability spaces hypotheses:** ADEs live in low-probability yet dense spaces in the data manifold that are not well represented in the training samples
   ▸ Szegedy et al. (2013)

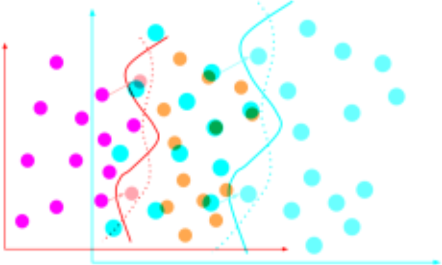


Figure: Binary classification example (dark blue vs. green dots). Dotted line represents the true decision boundary, bold line the trained one. Low probability space close to decision boundary allow for adversarial examples (turquoise dot).

WHY DO ADE EXIST?

Non-exhaustive list of hypotheses:

2. Linearity hypotheses (most popular):

Adversarial examples are omnipresent in the data manifold

$\rightsquigarrow$ occur, because commonly used models often show linear behavior

$\rightsquigarrow$ small changes of ϵ in every feature cause a change of $\epsilon \|\boldsymbol{\theta}\|_1$ in prediction

▸ Goodfellow et al. (2014)

Example: linear model

Original: $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\theta}$

Small changes: $f(\mathbf{x}+\epsilon) = (\mathbf{x}+\epsilon)^T\boldsymbol{\theta}$

Difference: $f(\mathbf{x}+\epsilon) - f(\mathbf{x}) = \epsilon \cdot \boldsymbol{\theta}$

Interpretable Machine Learning – 8 / 14

WHY DO ADE EXIST?

Non-exhaustive list of hypotheses:

3. The boundary tilting hypothesis: Linearity is neither necessary nor sufficient to explain ADEs

⇝ ADEs mostly result from overfitting the sampled manifold ▸ Tanay and Griffin (2016)

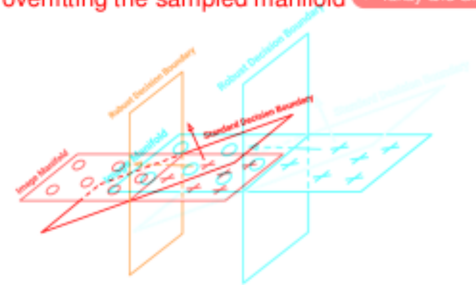
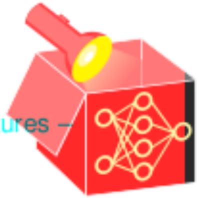


Figure: Linear binary classification example. Due to overfitting the decision boundary (red) is close to the manifold of the training data. Techniques like regularization could help to make the decision boundary more robust (green). ▸ Kim et al. (2019)

WHY DO ADE EXIST?



Non-exhaustive list of hypotheses:

4. **Human-centric hypotheses:** ML models make use of predictive but non-robust features – meaning they are highly correlated with the prediction target, but not used by humans ▶ Ilyas et al. (2019)

Interpretable Machine Learning – 8 / 14

WAYS TO GENERATE ADE

Different ways for constructing ADEs: There exist various ways in the literature to generate ADEs for a given model in feasible time

- Formulate the search for ADEs as an **optimization problem**, e.g.

$$\underset{\mathbf{x}' \in \mathcal{X}}{\text{argmin}} \underbrace{\|\mathbf{x} - \mathbf{x}'\|_{\mathcal{X}}}_{\text{minmize}} - \lambda \underbrace{\|f(\mathbf{x}') - y'\|_{\mathcal{Y}}}_{\text{maximize}}$$

- Use **sensitivity analysis** to identify features that influence the target class
- Train a generative adversarial network (GAN) ▸ Goodfellow et al. (2014)

Moreover, depending on the attacker's model access, we can distinguish between

- **Full-access attacks**: the attacker has full access to the internals of the model
- **Black-box attacks**: the attacker can only query the model on some inputs and receives the model's outputs

Interpretable Machine Learning – 9 / 14

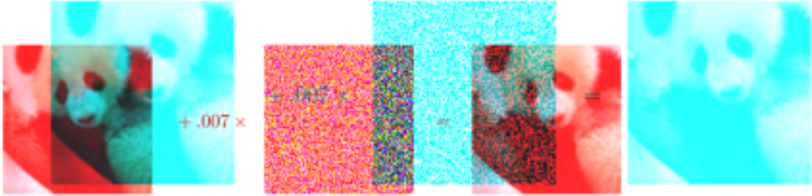
FAST GRADIENT SIGN METHOD (FGSM)

Goodfellow et al. (2015)

- FGSM is based on the linearity hypothesis
- FGSM finds ADEs from:

$$a_{\mathbf{x}} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y_{\mathbf{x},true}))$$

where $\text{sign}(\nabla_{\mathbf{x}} J(\theta, x, y_{\mathbf{x},true}))$ describes the component-wise signum of the gradient of cost function J in $\mathbf{x}$ with true label $y_{\mathbf{x},true}$



NOTES ON FGSM ▸ Goodfellow et al. (2015) ▸ Lyu et al. 2015

- FGSM works particularly well for linear(-like) models in high-dimensional spaces, e.g., LSTMs, logistic regressions or CNNs with ReLU activations
- Not every $\mathbf{a_x}$ generated by FGSM is an ADE, especially if ϵ is too small
- FGSM attacks can be also generated without model access by approximating the gradient, e.g. with finite difference methods
- The notion of similarity in FGSM is based on $\|\cdot\|_\infty$ $\rightsquigarrow$ there are generalizations of FGSM to other norms

Interpretable Machine Learning – 11 / 14

BLACK-BOX ATTACKS WITH SURROGATES

▸ Papernot et al. (2016)

- So far, we assumed full access to the predictive model
- Black-box attacks only assume query-access
- Large risk of attacks since often one can query predictive models many times

1. Query the model you aim to attack as often as allowed on data similar to the training data
2. Use the labeled data you received to train a surrogate model
3. Generate ADEs for the surrogate model
4. Use these ADEs to attack the original model

⇝ Known as the **transferability** of ADEs.

BLACK-BOX ATTACKS WITH SURROGATES ▸ Papernot et al. (2016)
- So far, we assumed full access to the predictive model
- Black-box attacks only assume query-access
- Large risk of attacks since often one can query predictive models many times
1. Query the model you aim to attack as often as allowed on data similar to the training data
2. Use the labeled data you received to train a surrogate model
3. Generate ADEs for the surrogate model
4. Use these ADEs to attack the original model
⇝ Known as the **transferability** of ADEs.

Interpretable Machine Learning – 12 / 14

DEFENSES AGAINST ADE

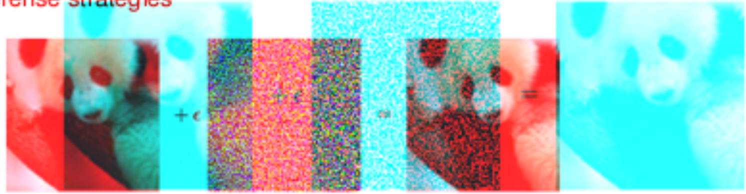
There are several ways to protect your network against such attacks – we distinguish between two broad types of defenses, differing in the position in which they act

- **Guards** act on the inputs a model receives
  - **Detect anomalies:** e.g., statistical testing, or discriminator networks from GANs
  - **Conduct transformations** on inputs (e.g. PCA)
- **Defense by design** act on the model itself
  - **Adversarial training:** train model on adversarials
  - **Architectural defenses:** e.g., removing low predictive features from the model

Interpretable Machine Learning – 13 / 14

SUMMARY

- ADEs are not explanations themselves but are conceptually connected to them
- ADEs can be generated in diverse settings $\rightsquigarrow$ crucial modeling decisions are the distance measure, the local environment, and the target level (model or process)
- There are various hypotheses on the existence of ADEs which also motivate different defense strategies



▸ Goodfellow et al. (2017)