AI4EO

Special Topics in Applications (AIL861)
Artificial Intelligence for Earth Observation
Lecture 16

Instructor: Sudipan Saha

Explainability

- ML/DL models: too many parameters
- Learned representations are complex
- Human-level understanding is desired

Some Keywords

- Trust
- Causality
- Transferability
- Fairness

More Terms

interpretable, implying some sense of understanding how the technology works

explainable, implying that a wider range of users can understand why or how a conclusion was reached

transparent, implying some level of accessibility to the data or algorithm;

justifiable, implying there is an understanding of the case in support of a particular outcome;

contestable, implying users have the information they need to argue against a decision or classification..

https://ec.europa.eu/futurium/en/system/files/ged/ai-and-interpretability-policy-briefing_creative_commons.pdf

Goals of Explainable AI

- Finding most important input features

- Find human-understandable concepts

- Discover new insights

- Identify model’s flaws

Explainable Models: Some Examples

- Local or individual predictions: LIME, SHAP
- Global or entire model: Partial dependence plots (PDP), aggregate LIME/SHAP

LIME

- Local Interpretable Model-agnostic Explanations (proposed in - Why Should I Trust You? Explaining the Predictions of Any Classifier)

- Explains the predictions of any classifier by learning an interpretable model locally around the prediction.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \;\; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

LIME

- Perturb the input/image (a possible perturbation can be to gray some super pixels)

- See how the predictions change

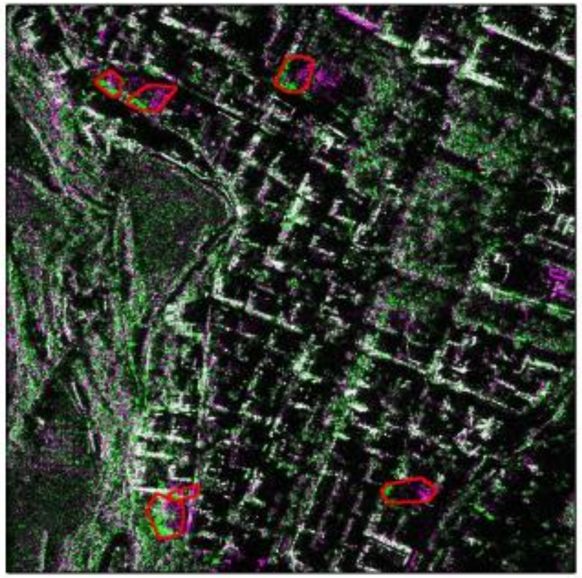
- By combining the perturbed instances, we can identify the region with highest weight (as explanation)

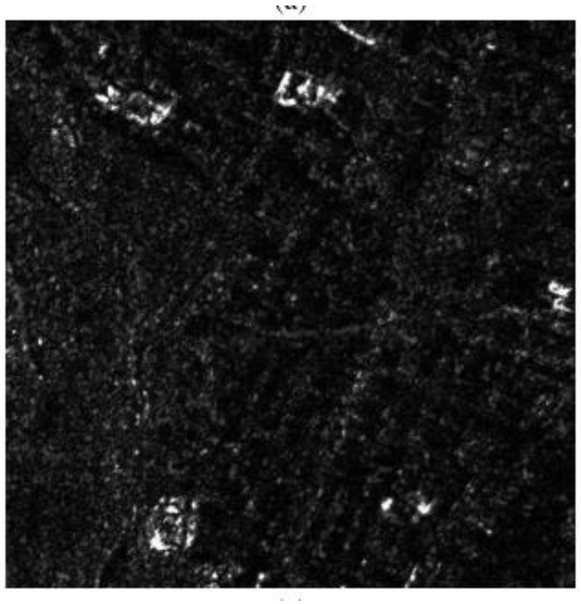
PDP

- It shows the marginal effect one or two features have on the predicted outcome of a machine learning model. A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic or more complex.

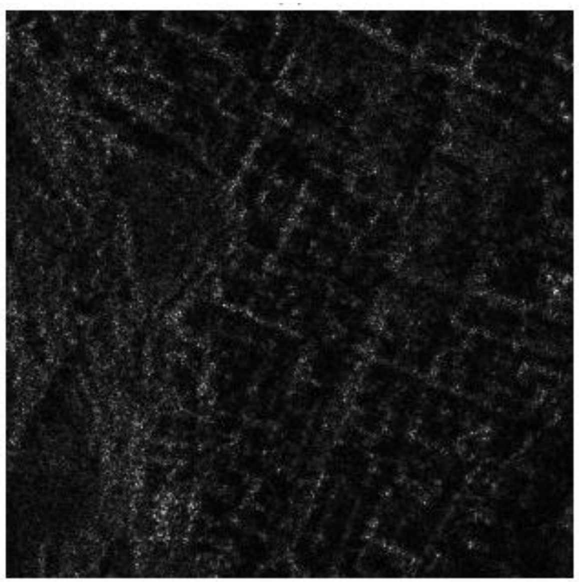
- See how the predictions change

Variance-Based Feature Selection for CD





One of the top features



One of the bottom features