Considerations

Surrogate fidelity

How can we better evaluate the surrogate's fidelity to the black box?

Surrogate fidelity

If black box makes a *continuous prediction* (including probability):

- Use continuous metrics (e.g., R^2, distance metrics).

We are not interested in the "classification" power of the surrogate, or how well the surrogate approximates the target.

Instead, we are interested in the **fidelity** of the surrogate, that is, how well it approximates the predictions of the black box.

Surrogate fidelity

If the black box makes a discrete prediction:

➔ We would use categorical metrics (e.g., accuracy, f1)

Advantages of surrogates

Flexibility

- We can use any interpretable model.
- We can explain any black box model.

Limitations

- We can’t (ultimately) be sure that the surrogate represent well the black box.

- If we chose a black box, probably we thought that a white box model was not appropriate. So using a white box to explain the black box is a bit contradictive.

Limitations

- Black box models have complex separation boundaries.
- White box models have simpler separation boundaries.

- Hard for a white box to approximate well a black box throughout the entire data space.

Alternatives

LIME: using surrogates for local / instance explanations (next section).

Alternatives

Surrogate within surrogate:

- Use a surrogate to split the data space into “explainable subsets”.
- Then fit a surrogate to explain each subset.

Surrogate examples

When the use of surrogates is worth it:

- To explain clustering models.
- To explain black box classifiers like one class support vector machines or isolation forests.

Train In Data

THANK YOU

www.trainindata.com