Generative Random Forests

Cassio de Campos

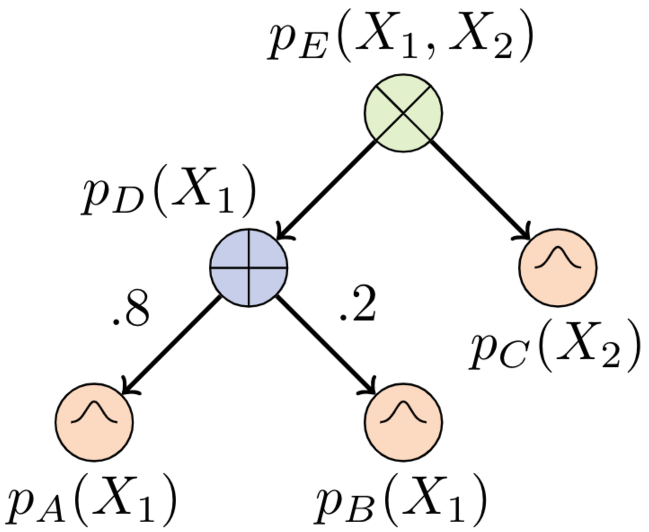
SIPTA Summer School 2022

Bristol

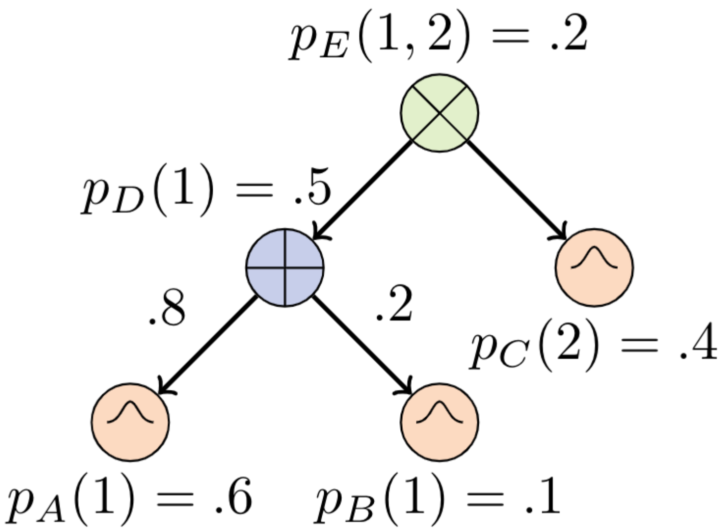
https://github.com/AlCorreia/GeFs

Probabilistic Circuits with continuous variables

Example



Evaluate at
$X_1 = 1$ and $X_2 = 2$

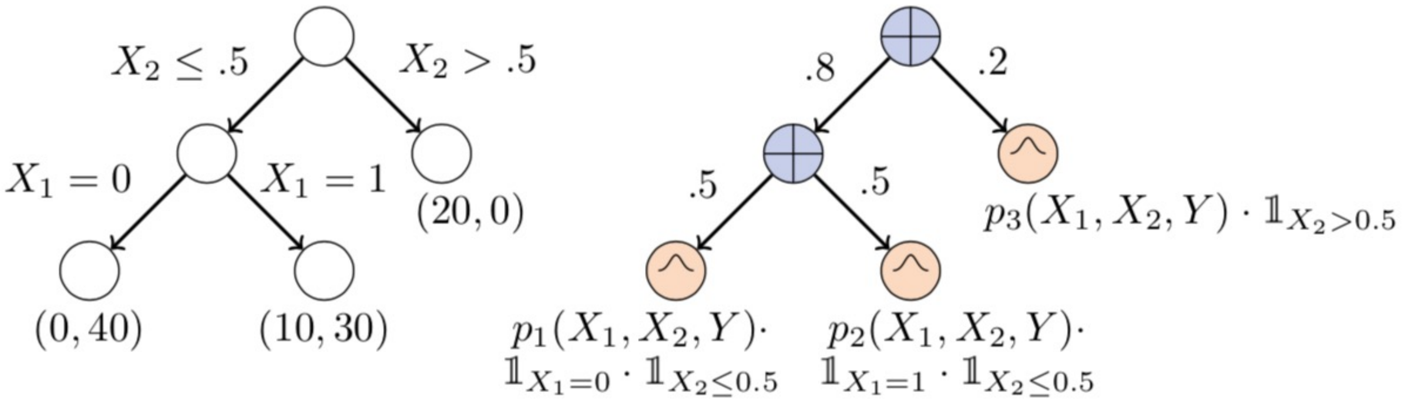


$$p_D(X_1) = 0.8 p_A(X_1) + 0.2 p_B(X_1)$$
$$p_E(X_1, X_2) = p_D(X_1) p_C(X_2) = \big(0.8 p_A(X_1) + 0.2 p_B(X_1)\big) p_C(X_2)$$

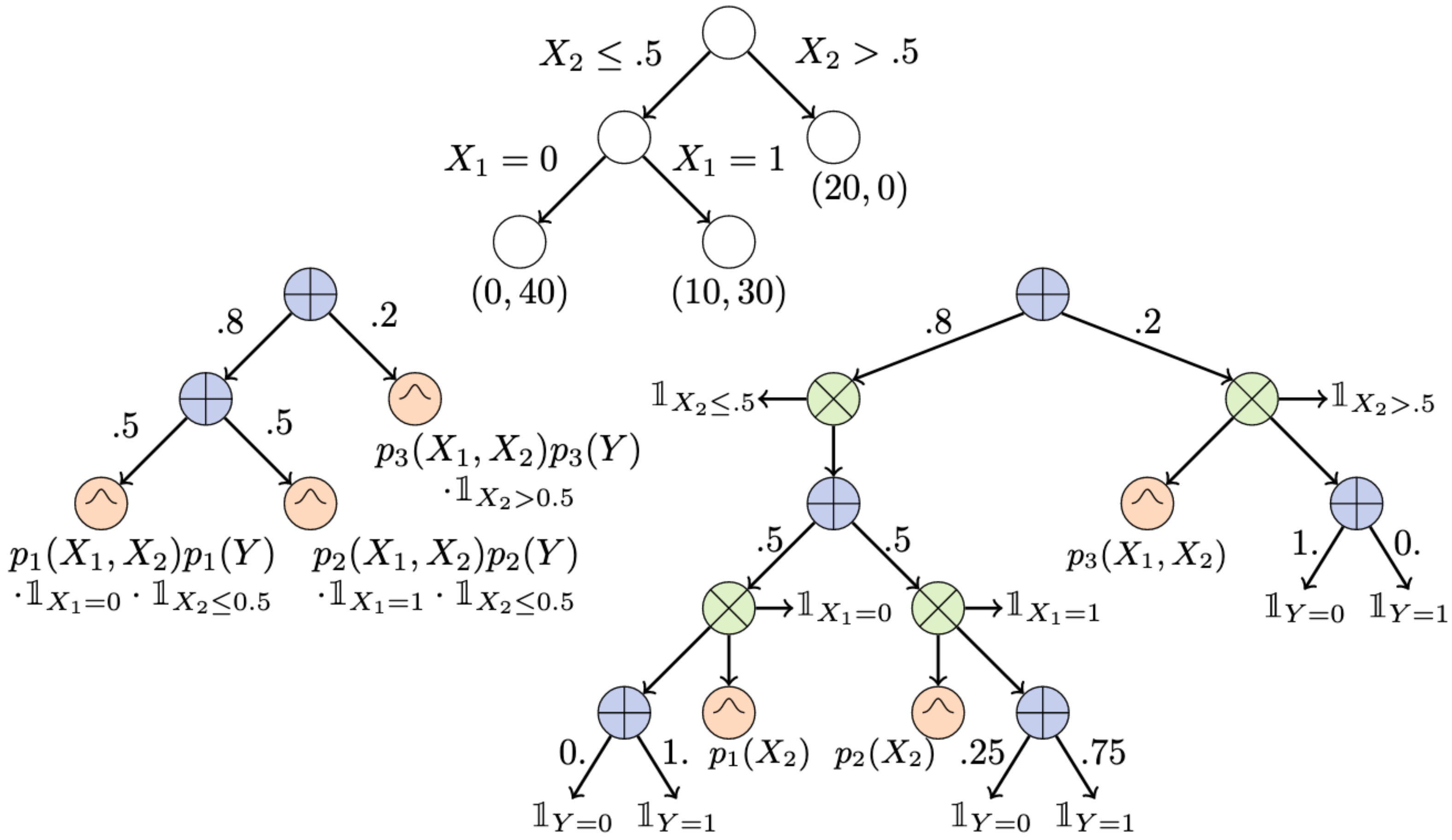
Generative Decision Trees (GeDTs)

- Representation of Decision Trees as Probabilistic Circuit

- Convert each internal node to a sum node
  - Weights are given by the mass of each children
- Convert each leaf into a distribution node
  - Fit a density over the instances in each leaf



Generative Decision Trees (GeDTs)



GeDTs to Generative Forests (GeFs)

- Ensemble of GeDTs just as RFs are ensembles of DTs

- Aggregation can be done by voting, averaging, etc, just as in RFs

We now can compute any marginals and conditional expectations!

We have information about the whole joint distribution $p(Y, \boldsymbol{X})$ (and not only $p(Y|\boldsymbol{X})$)

Outlier detection

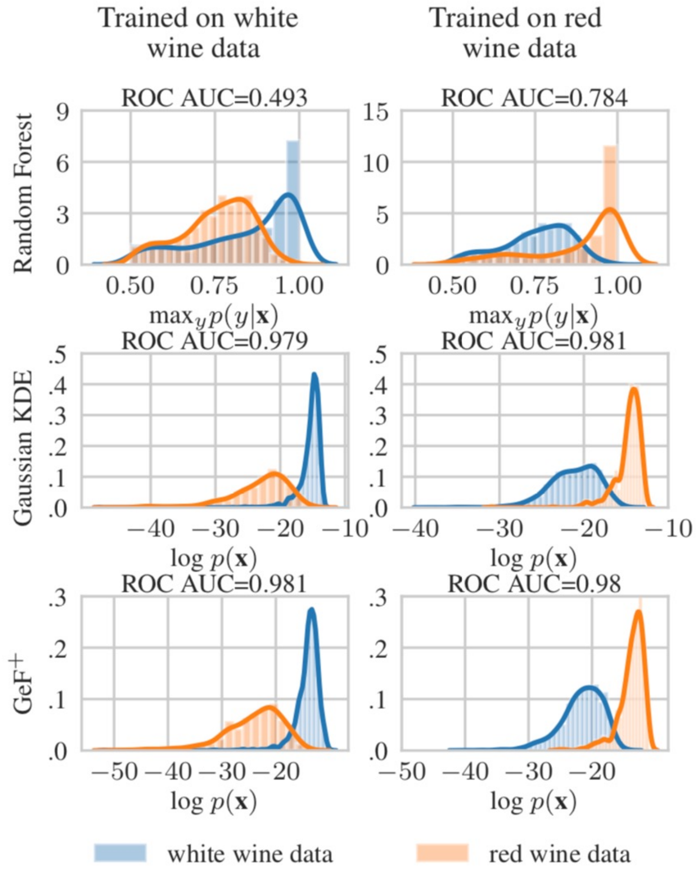
Generative Models are Natural Outlier Detectors

We detect out-of-domain samples simply by monitoring $p(\boldsymbol{X})$.

$$p(\boldsymbol{x}) = \sum_{y} p(y, \boldsymbol{x})$$

This comes at no extra cost, since Generative Forests perform classification over the joint, and all terms $p(y, \boldsymbol{x})$ are already computed in the context of classification.

$$\hat{y} = \mathrm{argmax}_y \, p(y, \boldsymbol{x})$$



Inference with MAR Missing Values

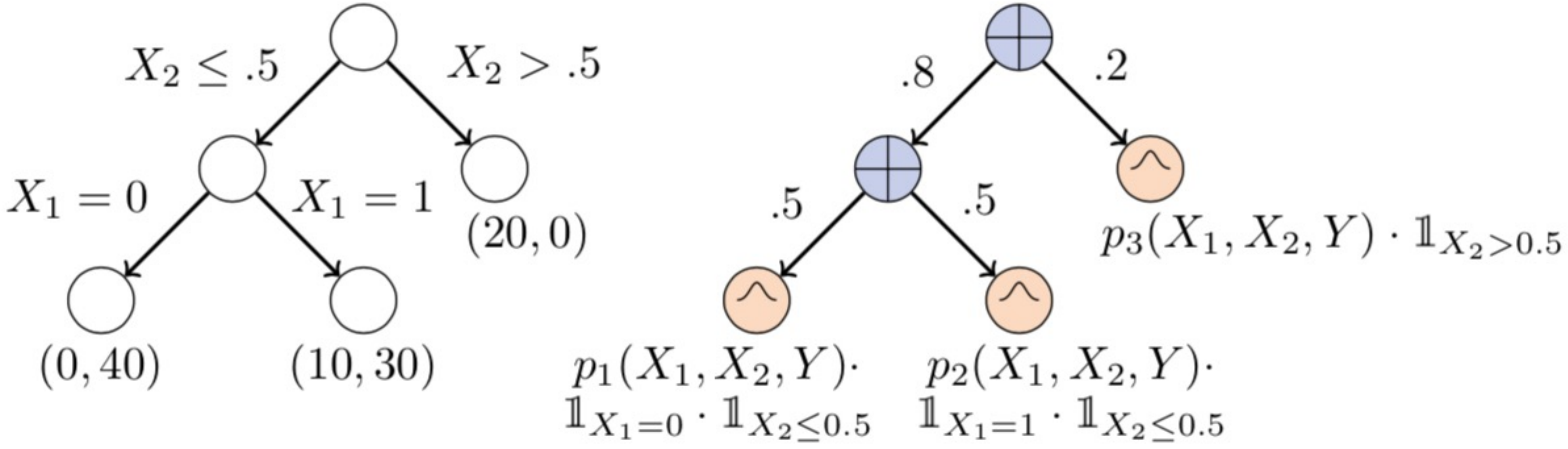
Marginalise the non-observed features $X_{\neg o}$

$$P(Y|X_o) = \frac{\int_{x_{\neg o}} P(Y, X_o, x_{\neg o}) dx_{\neg o}}{\sum_y \int_{x_{\neg o}} P(Y, X_o, x_{\neg o}) dx_{\neg o}}$$

Marginalisation with Generative Forests is tractable!

We can show this classifier is Bayes-consistent for any pattern of missing values!

What if X_2 is missing? Imputation? Surrogate split?

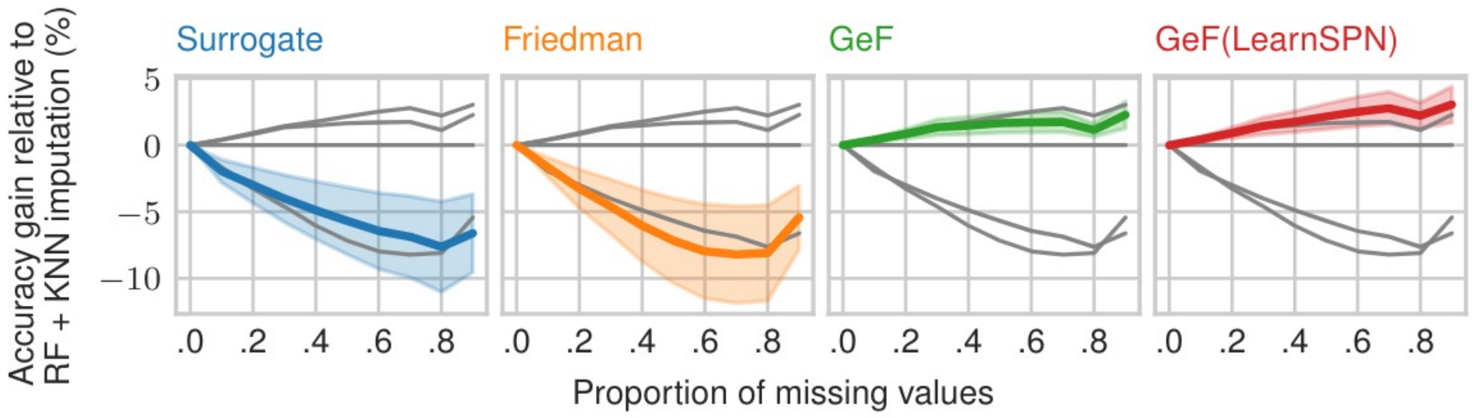


Missing at Random: Just compute with the marginalised variable – all works fine

Missing not at random: solve an optimisation similar to that of MAP/MPE, or https://link.springer.com/chapter/10.1007/978-3-030-86772-0_21 ask Alessandro about it

Inference with MAR Missing Values

Some Experimental Results



Average (across 21 datasets) accuracy gain relative to RFs (100 trees) plus KNN imputation against percent of missing values. Confidence intervals (95%) are also computed across the datasets.

Robust Classification (Attack on Parameters)

Sensitivity analysis

Perturb the model parameters until the predicted class changes.

ϵ-contamination of a vector of parameters $\boldsymbol{w}$

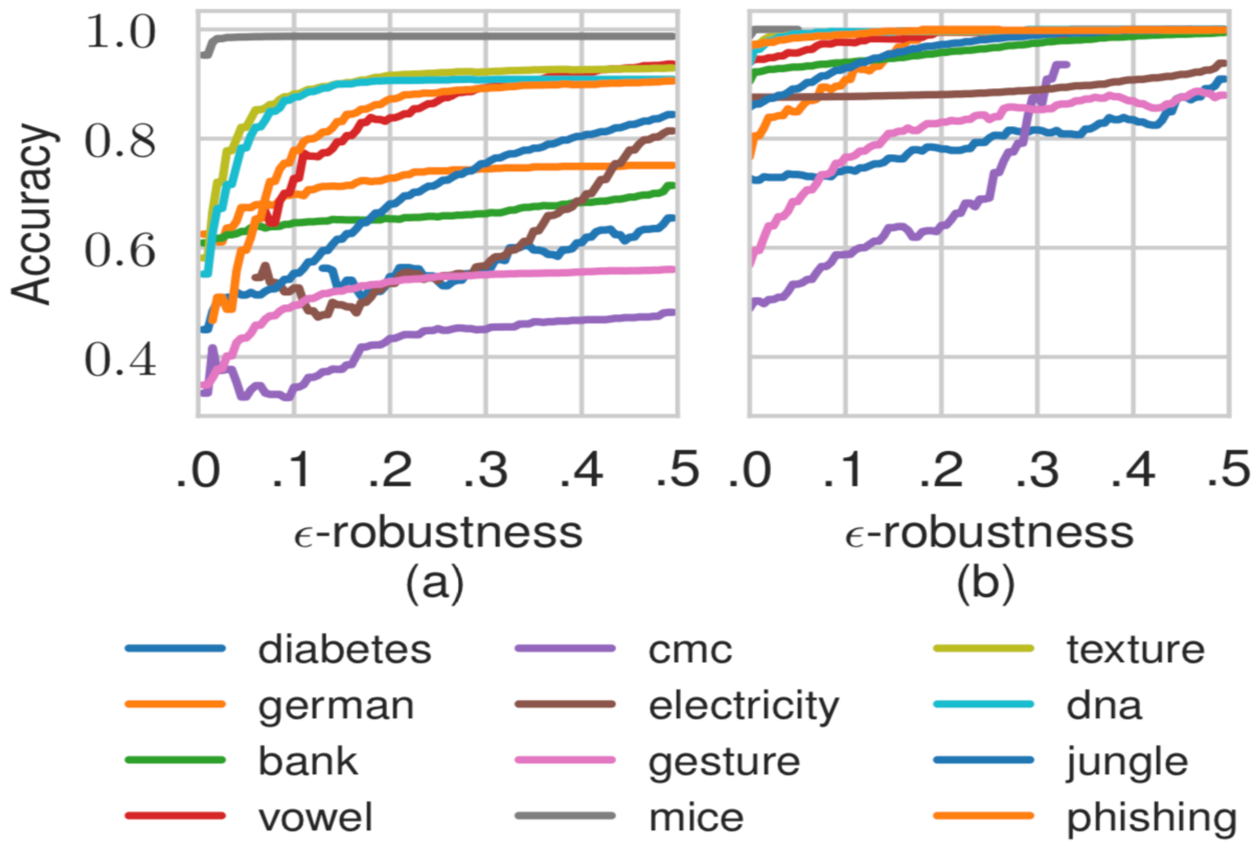
$$C_{\boldsymbol{w},\epsilon} = \{(1-\epsilon)\boldsymbol{w} + \epsilon\boldsymbol{v}: v_j \geq 0, \sum v_j = 1\}$$

ϵ-robustness

The largest ϵ for which all parameterisations in $C_{w,\epsilon}$ yield the same classification.

$$\forall y' \neq y: \max_{\boldsymbol{w}\in C_{\boldsymbol{w},\epsilon}} \mathbb{E}_{\boldsymbol{w}}\left[\mathbb{I}(Y = y') - \mathbb{I}(Y = y) \mid \boldsymbol{x}\right] < 0$$

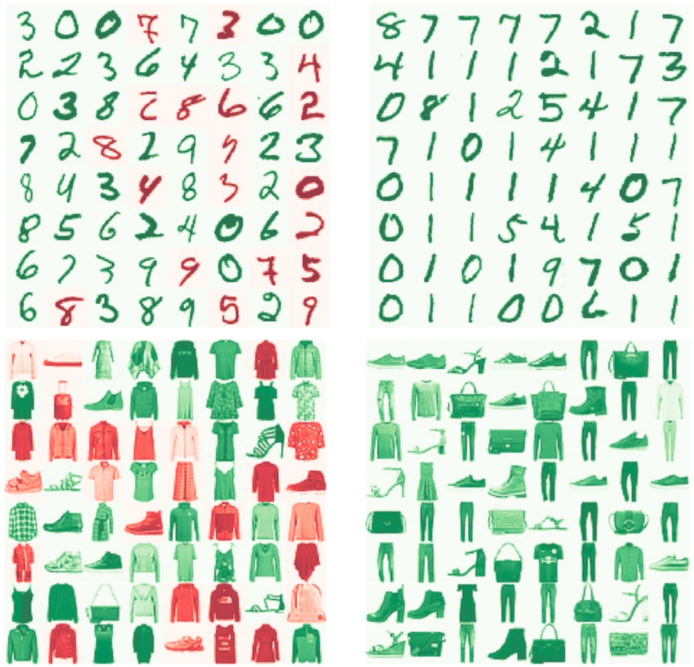
Robust Classification: ϵ-robustness correlates to accuracy



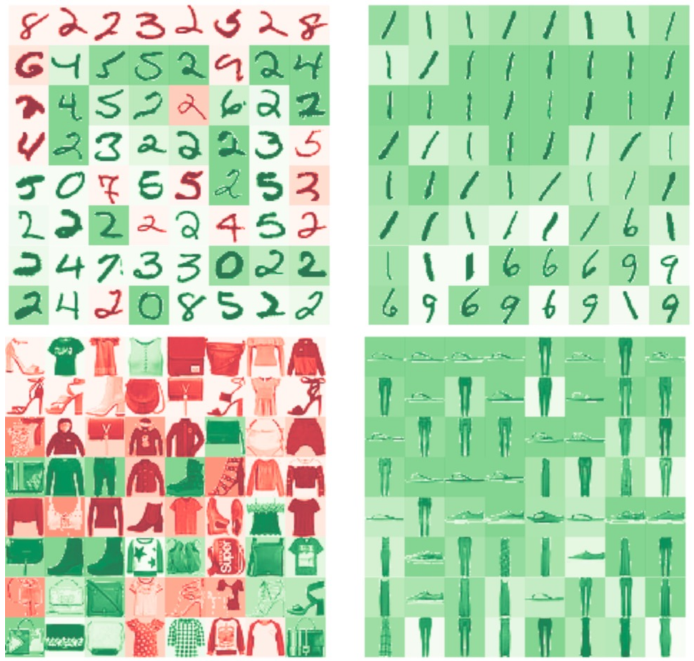
Accuracy of predictions with ϵ-robustness (a) below and (b) above different thresholds for 12 OpenML datasets.

Robust Classification

ϵ-robustness differs substantially from $p(\boldsymbol{x})$

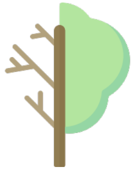


Samples from (Fashion-)Mnist datasets with lowest (left) and highest (right) ϵ-robustness in the test set.

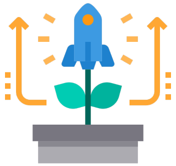


Samples from (Fashion-)Mnist datasets with lowest (left) and highest (right) $p(\boldsymbol{x})$ in the test set.

In a nutshell



- Extension of Random Forests to a full generative model.
  - Discriminative structure learning.
  - Model a joint distribution $p(Y, \boldsymbol{X})$ instead of a function $f: \boldsymbol{x} \mapsto y$.



- Enhance Random Forests
  - Outlier detection
  - Robust classification
  - Inference with missing values

https://github.com/AlCorreia/GeFs

Beyond Generative Forests: Credal Probabilistic Circuits

- So far only certain queries are known to be tractable
- Circuits are less powerful than hybrid Bayesian nets
  - There is need for further development to reach quality of other models such as Variational Auto-Encoders
- New theoretical properties allowing additional credal inferences

- Soft evidence
- Multi-label classification
- Strengthening relations between credal nets and credal circuits