

Interpretable Machine Learning

Ante-hoc Methods for Neural Networks



Learning goals

- Interpretability by sparsity
- Regularisation for interpretability
- Sequential feature selection

MOTIVATION

- Post-hoc methods do not always give you the entire picture
- Post-hoc methods are not always accurate
 - An explanation that is 10% inaccurate leads to lack of trust in the ML model
 - Hard to measure the accuracy of post-hoc methods
- Wherever possible use models that are interpretable-by-design



PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0049-z>

nature
machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

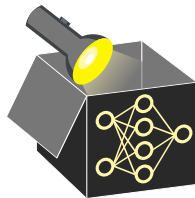
Cynthia Rudin



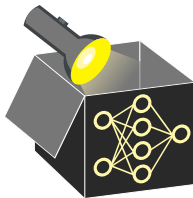
Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

SIMPLER MODELS

- Models that have an understandable decision-making process
- Models that have a smaller set of parameters or weights
 - Examples: Linear models, GAMs
- Models that have human-understandable decision structure
 - Examples: decision trees, random forests
- Models that have sparsity or only a few set of parameters or features that matter
 - Example: 1% of a large feature space, 1-hot encodings in language tasks



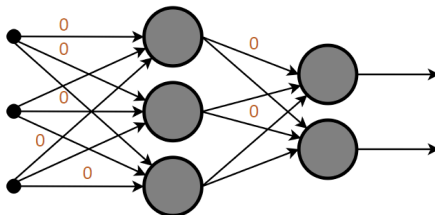
INTERPRETABLE BY DESIGN MODELS - SPARSE MODELS



- Models that have explicitly enforce sparsity
 - through regularisation
 - through feature selection
- Sparsity through regularisation
 - E.g. L0, L1 regularisation
- Sparsity through feature selection
 - select a subset of impacting features for the prediction task

REGULARISATION IN NEURAL NETWORKS

- L0 norm is the number of non-zero parameters — setting weights to 0
- L1 sparsity — sum of the weights should be small



L1 REGULARISATION

- Optimising using L0 regularisation is hard
- L1 regularisation in neural networks can be achieved by gradient-based optimisation
- Degree of regularisation is a user-controllable parameter

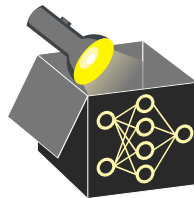


$$\hat{\mathcal{L}}(W) = \alpha \|W\|_1 + \mathcal{L}(W)$$
$$\nabla_w \hat{\mathcal{L}}(W) = \alpha \text{sign}(W) + \nabla_w \mathcal{L}(W)$$

FEATURE SELECTION

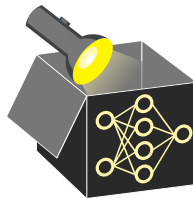
“Select a smaller features space which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results”

- Wrapper methods - Treat the model as a blackbox
 - Filter methods
 - Embedded methods
 - Other methods
-
- Smaller feature space: subset of features, an embedded hyperspace



SEQUENTIAL FEATURE SELECTION

- Number of feature subsets is 2^N
- How do we reduce the computational complexity of checking each subset ?
 - Sequentially choose the most promising feature at each iteration
- Selection Set $S = \{\}$, All features $N = \{f_1, f_2, \dots, F_n\}$
- In each iteration
 - compute utility of f - train a model with $S \cup \{f\}$ and measure validation perf.
 - terminate loop if no improvement of utility and return S
 - choose f in N/S that has max utility and add f to S



FEATURE SELECTION

- What are the shortcomings of sequential feature selection ?
 - Greedy might not be optimal
 - Global feature selection method
- How do we improve the greedy solution ?
 - Allow for backtracking, branch-and-bound
 - Use genetic algorithms GA, swarm optimisation
- How do we choose a local feature selection method ?
 - instance-wise feature selection methods

