Supervised and Interpretable Machine Learning in Medicine

Nataliya Sokolovska

Sorbonne University
Paris, France

Master 2 in Statistics
January, 14, 2020

Outline

Organisation

Medical Data and Some Applications

State-of-the-art Supervised ML Methods

Interpretable Models

Challenges in Interpretable Supervised Learning

Organisation

Organisation

Medical Data and Some Applications

State-of-the-art Supervised ML Methods

Interpretable Models

Challenges in Interpretable Supervised Learning

Organisation

10 (dense) lectures + 10 practical sessions (Python)

1. Supervised learning and interpretable models
2. Unsupervised and semi-supervised learning
3. Structured output prediction and feature selection
4. Text and medical text processing
5. Causal inference
6. Network reconstruction
7. Deep learning
8. Learning under budget
9. Kinetic data
10. Project defence

Project and notes

Final note:

- 30% TME
- 70% project

Project:

- 2 TME dedicated to the project
- 1 TME dedicated to its defence
- The goal: choose a data set of your interest (from the UCI ML repository), explore it by applying existing (supervised, unsupervised learning methods). Propose an algorithmic improvement, present your results (prepare slides, code, no report).

Organisation

Medical Data and Some Applications

State-of-the-art Supervised ML Methods

Interpretable Models

Challenges in Interpretable Supervised Learning

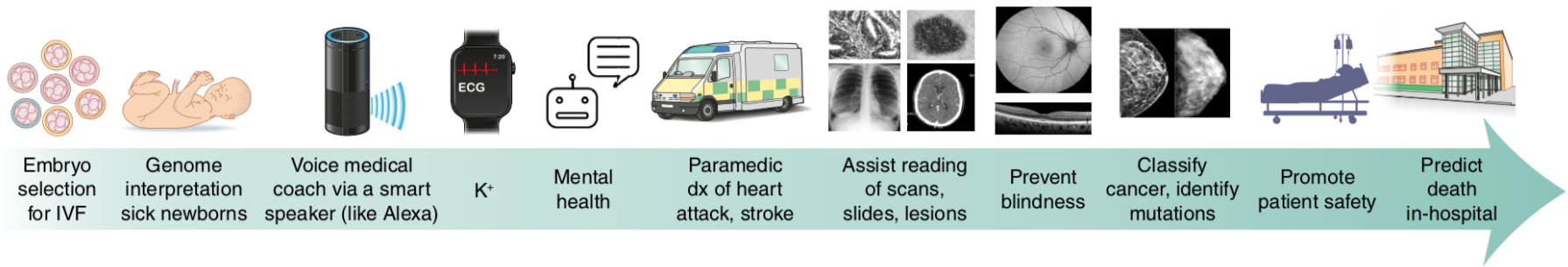
Medical Data

- Small number of observations (N)
- Big number of parameters p
- Data are noisy
- Missing data
- Batch effect (possible)
- In a real medical study, data are usually heterogeneous

Heterogeneous Medical Data and Data Integration

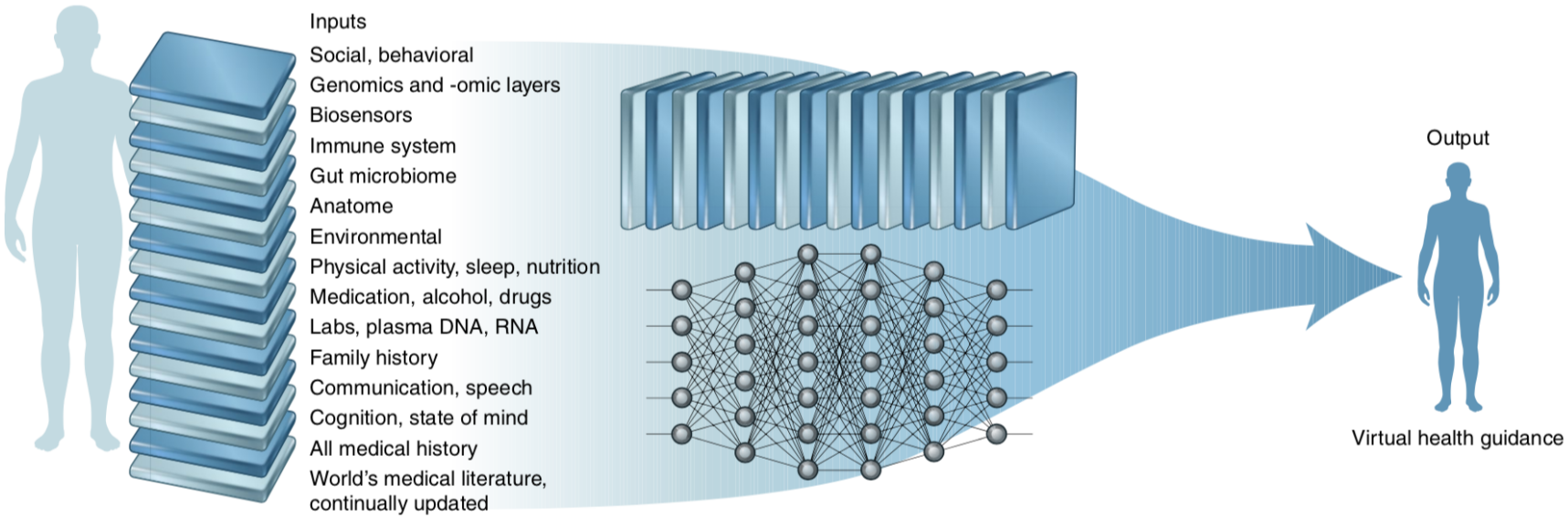
- Clinical data, alimentary patterns, nutritional habits
- Drugs taken, treatement
- “omics” data
  - lipidomics
  - transcriptomics
  - metagenomics
  - proteomics

Machine Learning Applications in Medicine



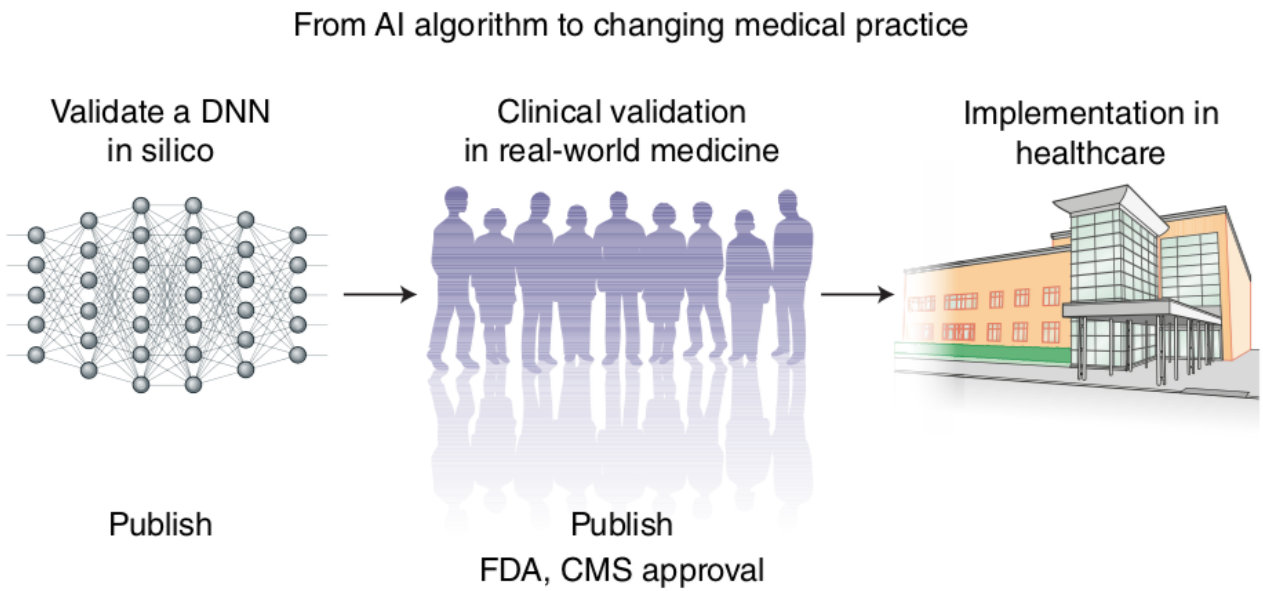
from *E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 2019.*

Multi-modal data inputs to provide individualized guidance



from *E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 2019.*

Machine Learning Studies and Medical Routines



from *E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 2019.*

Organisation

Medical Data and Some Applications

State-of-the-art Supervised ML Methods

Interpretable Models

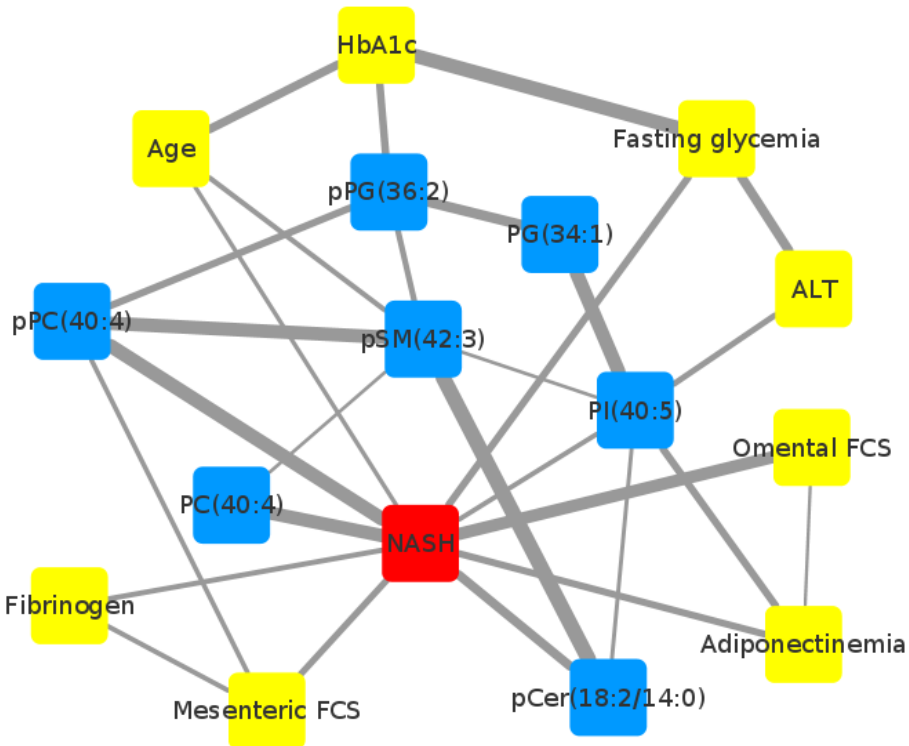
Challenges in Interpretable Supervised Learning

State-of-the-art Supervised ML methods

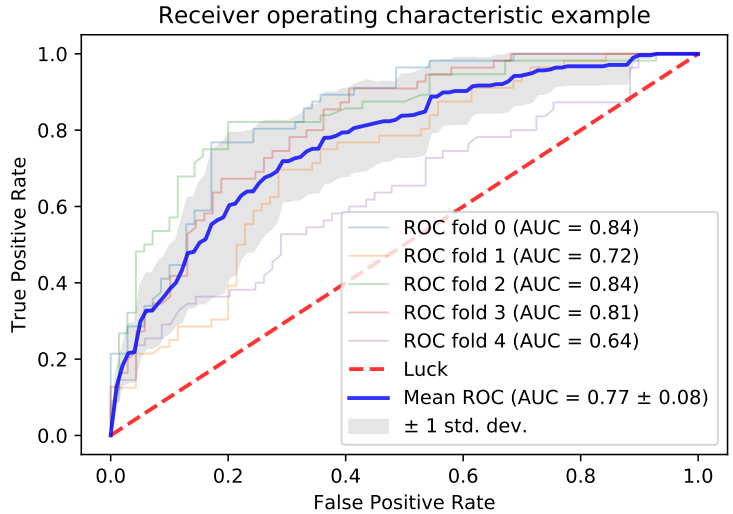
- Logistic regression
- Support Vector Machines
- Random Forests
- Boosting and Gradient Boosting
- Deep Learning

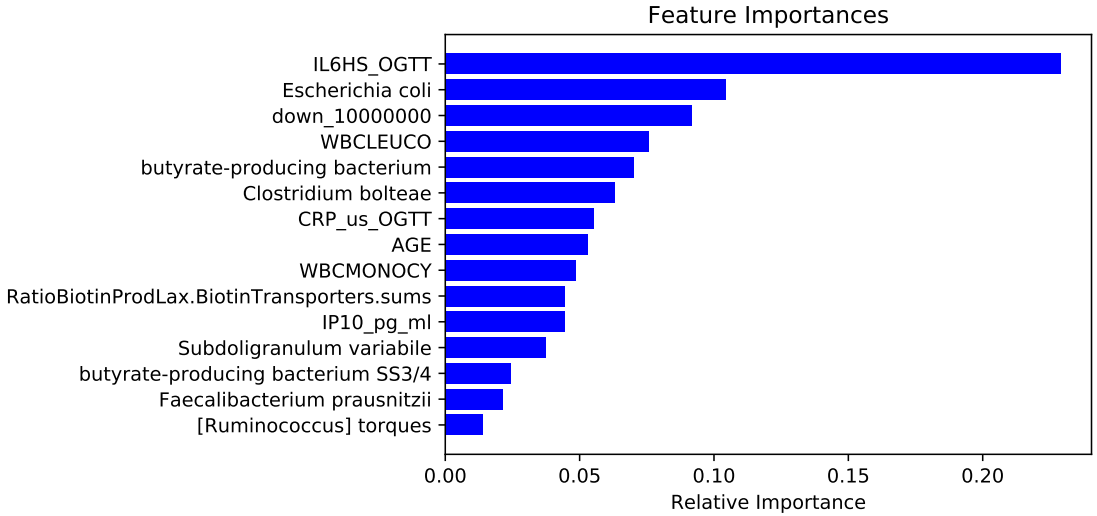
Support Vector Machines in Medicine

- K. Anjani, M. Lhomme, N. Sokolovska, C. Poitou, J.-L. Bouillot, P. Lesnik, P. Bedossa, A. Kontush, K. Clément, I. Dugail, J. Tordjman. *Circulating phospholipid profiling identifies portal contribution to NASH signature in obesity*, Journal of Hepatology, 2015.
- Find the best predictors for NASH
- Apply a sparse SVM, perform also feature selection
- Visualise the selected features by a Bayesian net



Random Forests in Medicine





Black box models

- *Black box* vs *interpretable* or explainable models
- Interpretability is not well defined
- Not interpretable by human experts
- Some models have some interpretable aspects
- A classical example of a black box model: neural (deep) networks (typically involve non-linearities and interactions between inputs, which means that not only is there no simple mapping from input to outputs, the effect of changing one input may dependent critically on the values of other inputs. This makes it very hard to mentally figure out what?s happening)

Organisation

Medical Data and Some Applications

State-of-the-art Supervised ML Methods

Interpretable Models

Challenges in Interpretable Supervised Learning

Explainable Decisions

- Fairness: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.
- Privacy: Ensuring that sensitive information in the data is protected.
- Reliability: or Robustness: Ensuring that small changes in the input do not lead to large changes in the prediction.
- Causality: Check that only causal relationships are picked up.
- Trust: It is easier for humans to trust a system that explains its decisions compared to a black box.

Fairness in Machine Learning

A hot topic in Machine Learning:

- Unintended discrimination arises naturally and frequently in the use of machine learning and algorithmic decision making
- The focus is on understanding and mitigating discrimination based on sensitive characteristics, such as, gender, race, religion, physical ability, and sexual orientation

Why it happens?

- A learning algorithm is designed to pick up statistical patterns in training data
- If the training data reflect existing social biases against a minority, the algorithm is likely to incorporate these biases.

Interpretable Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + \epsilon \tag{1}$$

- βs are the learned feature weights
- β_0 is the intercept, ϵ is the error the model makes (Gaussian distribution)
- Numerical feature: Increasing the numerical feature by one unit changes the estimated outcome by its weight. An example of a numerical feature is the size of a house
- We hope that there are not any strongly correlated features
- In the medical field, it is not only important to predict the clinical outcome of a patient, but also to quantify the influence of the drug and at the same time take sex, age, and other features into account in an interpretable way

Interpretable Logistic Regression

A linear model (above) does not output probabilities, but it treats the classes as numbers (0 and 1)

$$\log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d \tag{2}$$

- The outcomes are probabilities
- A change in a feature by one unit changes the odds ratio (multiplicative) by a factor of $\exp(\beta_j)$

Decision Trees

- Linear models fail is the relationship between classes and features is non-linear
- Tree based models split the data multiple times according to certain cutoff values in the features
- Each instance falls into exactly one leaf node
- Feature Importance: Go through all the splits for which the feature was used and measure how much it has reduced the variance or Gini index compared to the parent node (The sum of all importances is scaled to 100)
- Advantages: capturing interactions between features, good visualisation
- Disadvantages: unstable, the number of terminal nodes increases quickly with depth (difficult to interpret)

Motivation and Goals

- **Motivation**
  - **Simple** and **interpretable** models
- **A scoring system**
  - **sparse linear** model
  - based on **simple arithmetic operations**
  - has **few significant digits** (ideally integers)
  - can be **explained by human experts**
  - to be learned purely from data

Example: the DiaRem (Diabetes Prediction) Score

Variable	Thresholds	Score
Age	<40	0
	40–49	1
	50 – 59	2
	>60	3
Glycated hemoglobin	<6.5	0
	6.5 – 6.9	2
	7 – 8.9	4
	> 9	6
Insuline	No	0
	Yes	10
Other drugs	No	0
	Yes	3

Classify as Remission if sum of scores < 7
Classify as Non-remission if sum of scores ≥ 7

C. D. Still et al., Preoperative prediction of type 2 diabetes remission after Roux-en-Y gastric bypass surgery: a retrospective cohort study, 2013

The State-of-the-Art

Medical Scores (widely used)

- SAPS I, II, and III and APACHE I, II, III to assess intensive care units mortality risks
- $CHADS_2$ to assess the risk of stroke
- TIMI to estimate the risk of death of ischemic events

None of the existing medical scores was learned directly from data without any human manipulation.

State-of-the-Art Cont'd

Machine Learning point of view:

- Problems are formulated and solved as linear integer tasks
  - *B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. Machine Learning, 2015.*
- Bayesian optimisation is used to fit a model
  - *S. Ertekin and C. Rudin. A Bayesian approach to learning scoring systems. Big Data, 3(4), 2015.*
- Linear methods (regressions) using gradient-based optimisation, with rounded coefficients
  - *D. Golovin, D. Sculley, H. B. McMahan, and M. Young. Large-scale learning with less ram via randomization. In ICML, 2013.*

Automated Score Construction

1. Identification of related clinical variables

age	glycated hemoglobin	insuline	other drugs

Automated Score Construction

1. Identification of related clinical variables

age	glycated hemoglobin	insuline	other drugs

2. Meaningful thresholds for clinical variables

age				glycated hemoglobin				insuline		other drugs	
<40	40–49	50 – 59	>60	<6.5	6.5 – 6.9	7 – 8.9	> 9	yes	no	yes	no

Automated Score Construction

1. Identification of related clinical variables

age	glycated hemoglobin	insuline	other drugs

2. Meaningful thresholds for clinical variables

age				glycated hemoglobin				insuline		other drugs	
<40	40–49	50 – 59	>60	<6.5	6.5 – 6.9	7 – 8.9	> 9	yes	no	yes	no

3. Optimization of weights for sub-groups of the variables

age				glycated hemoglobin				insuline		other drugs	
<40	40–49	50 – 59	>60	<6.5	6.5 – 6.9	7 – 8.9	> 9	yes	no	yes	no
0	1	2	3	0	2	4	6	10	0	3	0

Automated Score Construction

1. Identification of related clinical variables

age	glycated hemoglobin	insuline	other drugs

2. Meaningful thresholds for clinical variables

age				glycated hemoglobin				insuline		other drugs	
<40	40–49	50 – 59	>60	<6.5	6.5 – 6.9	7 – 8.9	> 9	yes	no	yes	no

3. Optimization of weights for sub-groups of the variables

age				glycated hemoglobin				insuline		other drugs	
<40	40–49	50 – 59	>60	<6.5	6.5 – 6.9	7 – 8.9	> 9	yes	no	yes	no
0	1	2	3	0	2	4	6	10	0	3	0

4. Find an optimal separator between two classes

Classify as `Remission` if sum of scores < 7
Classify as `Non-remission` if sum of scores ≥ 7

Our team worked on:

- **Simultaneously** do: **binning** (a supervised discretization) and the **score learning** for the bins.
- The **Fused Lasso** (*R. Tibshirani et al., 2015*) shrinks similar variables to each other creating bins, and ordering them.
- In our approach: the **Fused Lasso creates categories and estimates the corresponding weights.**

The Linear Formulation

We minimise the hinge loss

$$\sum_{i=1}^{N} \ell(y_i, \theta \cdot \bar{x}_i + b) + \lambda \sum_{j=1}^{\bar{d}-1} |\theta_j - \theta_{j+1}|. \tag{3}$$

If we re-write the task as an optimisation problem, we obtain:

$$\min \Big(\sum_{i=1}^{N} \xi_i + \sum_{j=1}^{\bar{d}} \eta_j \Big), \text{ such that} \tag{4}$$

$$\text{for all } i, \; y_i(\theta \cdot \bar{x}_i + b) \geq 1 - \xi_i, \tag{5}$$

$$\text{for all } j, \; -\lambda\eta_j \leq \theta_j - \theta_{j+1} \leq \lambda\eta_j, \tag{6}$$

$$\xi_i \geq 0, \theta_i \in \mathbb{N} \text{ for all } i, \tag{7}$$

and we get $\bar{d} + 1 + N + (\bar{d} - 1)$ variables
$\theta_1, \ldots, \theta_{\bar{d}}, b, \xi_1, \ldots, \xi_N, \eta_1, \ldots, \eta_{\bar{d}-1}$.

The **Algorithm**: a Linear SVM Penalized by Fused Lasso for Score Learning

Input: a continuous matrix X $(N \times d)$, class vector Y
Output: weights associated with each (observed) value in X

for $j \in \{1, \ldots, d\}$ **do**

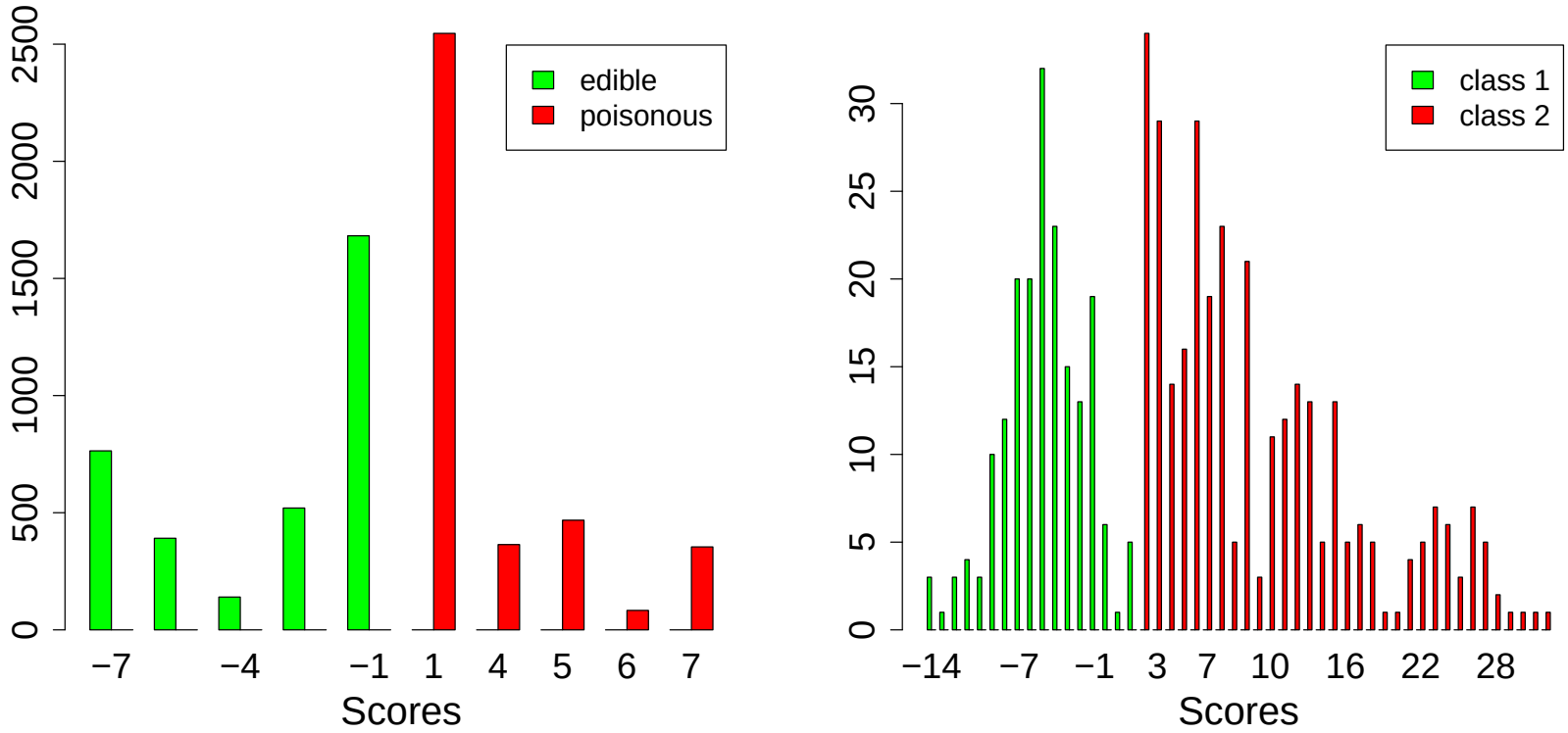
Reformulate X^j as a matrix $\bar{X}$ using one-hot-encoding

Solve discrete L1-SVM with integrity constraints on θ and fused-lasso penalty using $\bar{X}$ and Y

From the resulting θ, build a binning of the values of X^j, such that two contiguous values associated with equal weights are in the same bin

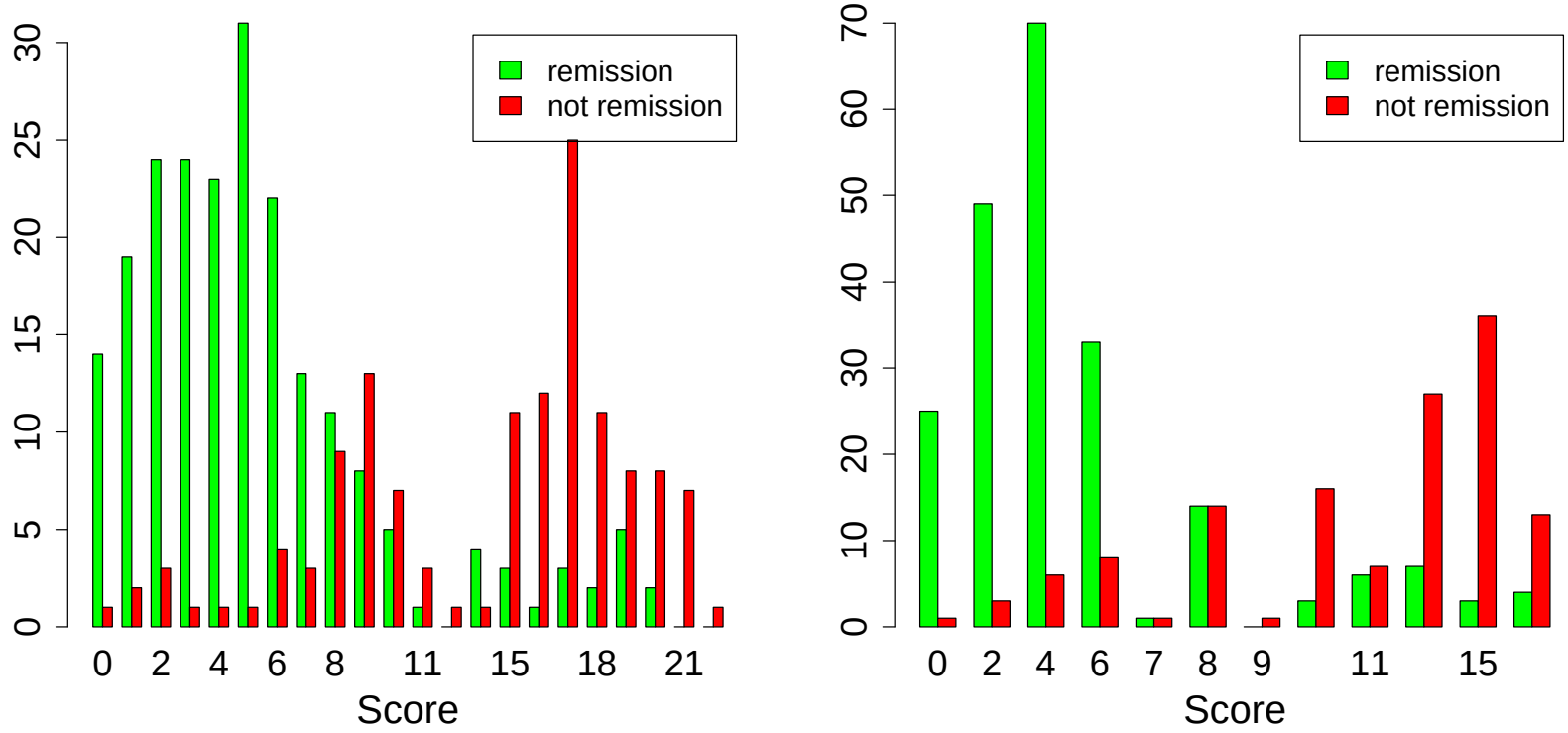
end for

The obtained **scores**: Mushrooms and Breast Cancer Data Sets



Distributions of the scores on the Mushrooms data (on the left), and on the Breast cancer data (on the right). On the horizontal axis: all possible scores in data sets. On the vertical axis: the number of observations with the corresponding score. The classes are quite well separated; the optimal separator value is 0.

Prediction of the Diabetes Remission



Distributions of patients according to the diabetes remission scores. On the left: scores obtained with the DiaRem score, on the right: a distribution based on the learned scoring system.

FCB

We define the problem of scoring systems learning as follows. We have a set of training examples $\{Z_i, Y_i\}_{i=1}^{N}$, where Z is the interval encoding of some matrix X, and Y is a class label. A score function is defined as $\langle \theta, Z \rangle$, where θ is a coefficient vector, and $\langle \cdot, \cdot \rangle$ is the scalar product. Given Z, and estimated weights θ, a score s_i for an observation Z_i is equal to $\langle \theta, Z_i \rangle$. A class can be predicted according to the conditional probability

$$p(y = 1|Z) = \frac{1}{1 + \exp(-\langle \theta, Z \rangle)}. \tag{8}$$

FCB cont'd

The problem is formulated as a feature selection task.
The proposed algorithm at each iteration finds an optimal model over all already added features, and adds a new feature, i.e., splits one of the existing bins into two bins, if this operation minimizes the empirical risk:

$$j, l, u, r = \underset{\text{for all } j,]l,u],r\in]l,u]}{\text{argmax}} \Big(\max(|(\nabla R)_{jlr}|, |(\nabla R)_{jru}|) \Big), \quad (9)$$

$$\theta = (\theta \cup \{\theta_{jlr}, \theta_{jru}\}) - \{\theta_{jlu}\}. \quad (10)$$

In a replacement step of the algorithm, the least important feature

$$j, l, u, q = \underset{\text{for all } j,]l,q],]q,u],q\in]l,u]}{\text{argmin}} \Big(|\theta_{jlq} - \theta_{jqu}| \Big), \quad (11)$$

$$\theta = (\theta \cup \{\theta_{jlu}\}) - \{\theta_{jlq}, \theta_{jqu}\}. \quad (12)$$

is removed from the model if this operation does not degrade the performance. In other words, one of the bins is merged with its neighbour.

FCB cont'd

Input: Training data $\{X_i, Y_i\}_{i=1}^N$, $\quad X : N \times p$
Output: Scoring model θ

Construct matrix Z from X according to eq. (3.1) // Initialize the bins

for all $j \in \{1, \ldots, p\}$

$\quad \theta_{j-\infty+\infty} = 0$ // Initialize the weights

end for

$\theta = \arg \min_{\text{supp}(\theta)} R(\theta)$ // Update the parameters

for $t = 1, \ldots, T$

$\quad j, l, u, r = \text{argmax}_{\text{for all } j,]l,u],r\in]l,u]} \Big(\max(|(\nabla R)_{jlr}|, |(\nabla R)_{jru}|) \Big)$, // Split (add) a variable and update the binning

$\quad \theta = (\theta \cup \{\theta_{jlr}, \theta_{jru}\}) - \{\theta_{jlu}\}.$

$\quad \theta = \arg \min_{\text{supp}(\theta)} R(\theta)$ // Update the parameters, update Z

$\quad$ **if** $t > K$

$\quad\quad j, l, u, q = \text{argmin}_{\text{for all } j,]l,q],]q,u],q\in]l,u]} \Big(|\theta_{jlq} - \theta_{jqu}| \Big)$, // Merge (delete) a variable and update the binning

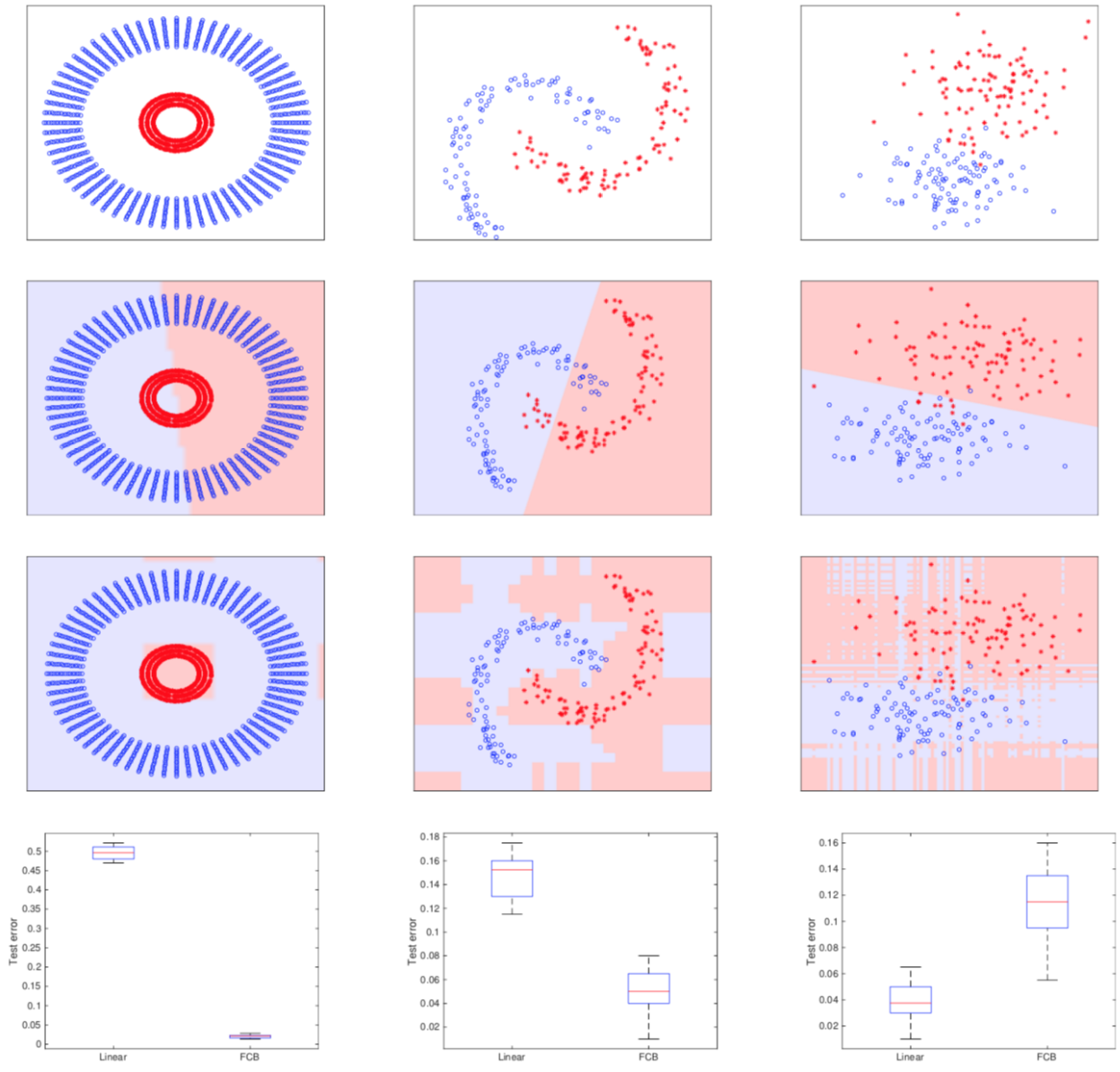
$\quad\quad \theta = (\theta \cup \{\theta_{jlu}\}) - \{\theta_{jlq}, \theta_{jqu}\}.$

$\quad\quad \theta = \arg \min_{\text{supp}(\theta)} R(\theta)$ // Update the parameters again, update Z

$\quad$ **end if**

end for

FCB cont'd


Test error
Linear
FCB
Test error
Linear
FCB
Test error
Linear
FCB

SLIM

Supersparse Linear Classification Models
B. Ustun, S. Traca, C. Rudin, *Supersparse Linear Integer Models for Interpretable Classification*, 2014

- L_1 norm is used to find a sparse solution
- L_0 norm is a an "ideal" penalty

$$\min_{\lambda} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y_i \lambda^T x_i \leq 0\} + C_0 \|\lambda\|_0 + \epsilon \|\lambda\|_1 \tag{13}$$

SLIM optimises accuracy and sparsity by minimising the $0-1$ loss the the L_0 norm.

SLIM cont'd

The task is presented and solved as an integer programming problem, and we use the Matlab implementation[1] provided by the SLIM authors. The training procedure relies on the IBM ILOG CPLEX Optimization Studio[2] which efficiently performs the constrained optimization. In particular, integrity constraints are added to the optimisation problem to obtain integer solutions.

[1] https://github.com/ustunb/slim-matlab

[2] http://www-03.ibm.com/software

Local interpretable model-agnostic explanations (LIME)

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016

- Surrogate models are trained to approximate the predictions of the underlying black box model
- Instead of training a global surrogate model, LIME focuses on training local surrogate models to explain why individual predictions were made.

The goal is to understand why the machine learning model made a certain prediction. LIME tests what happens to the predictions when you give variations of your data into the machine learning model. LIME generates a new dataset consisting of permuted samples and the corresponding predictions of the black box model. On this new dataset LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest.

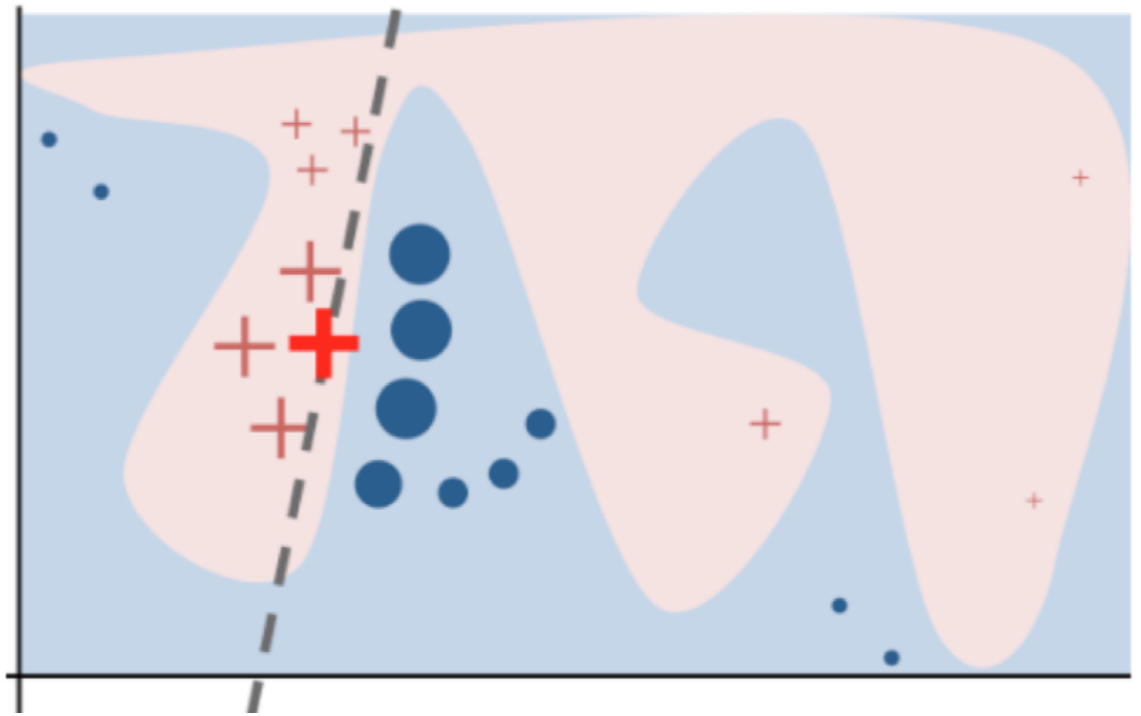
LIME cont'd

The recipe for training local surrogate models:

1. Select your instance of interest for which you want to have an explanation of its black box prediction.
2. Perturb your dataset and get the black box predictions for these new points.
3. Weight the new samples according to their proximity to the instance of interest.
4. Train a weighted, interpretable model on the dataset with the variations.
5. Explain the prediction by interpreting the local model.

LIME cont'd

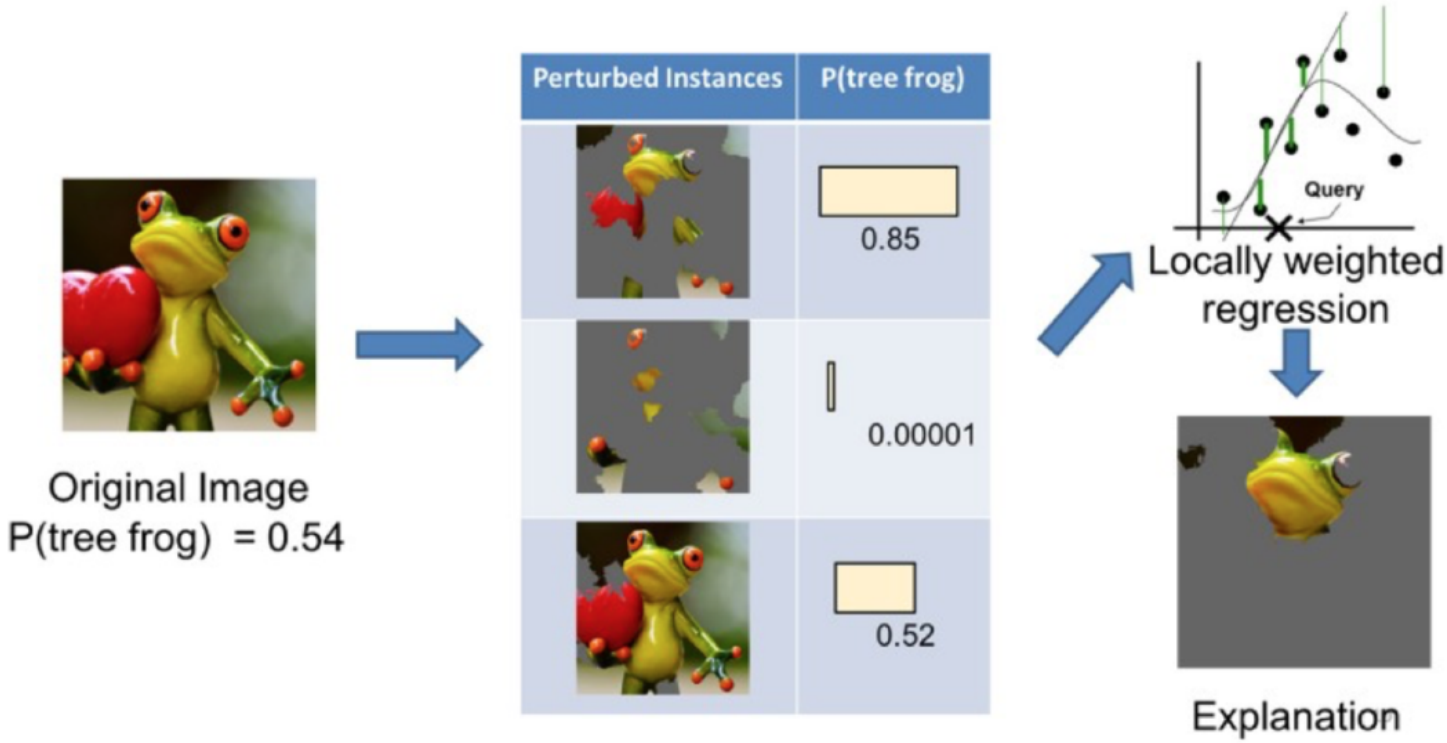
Learn an interpretable model (e.g., linear model) in the vicinity of the given instance.



M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135?1144. ACM, 2016.

LIME cont'd

Learn an interpretable model (e.g., linear model) in the vicinity of the given instance.

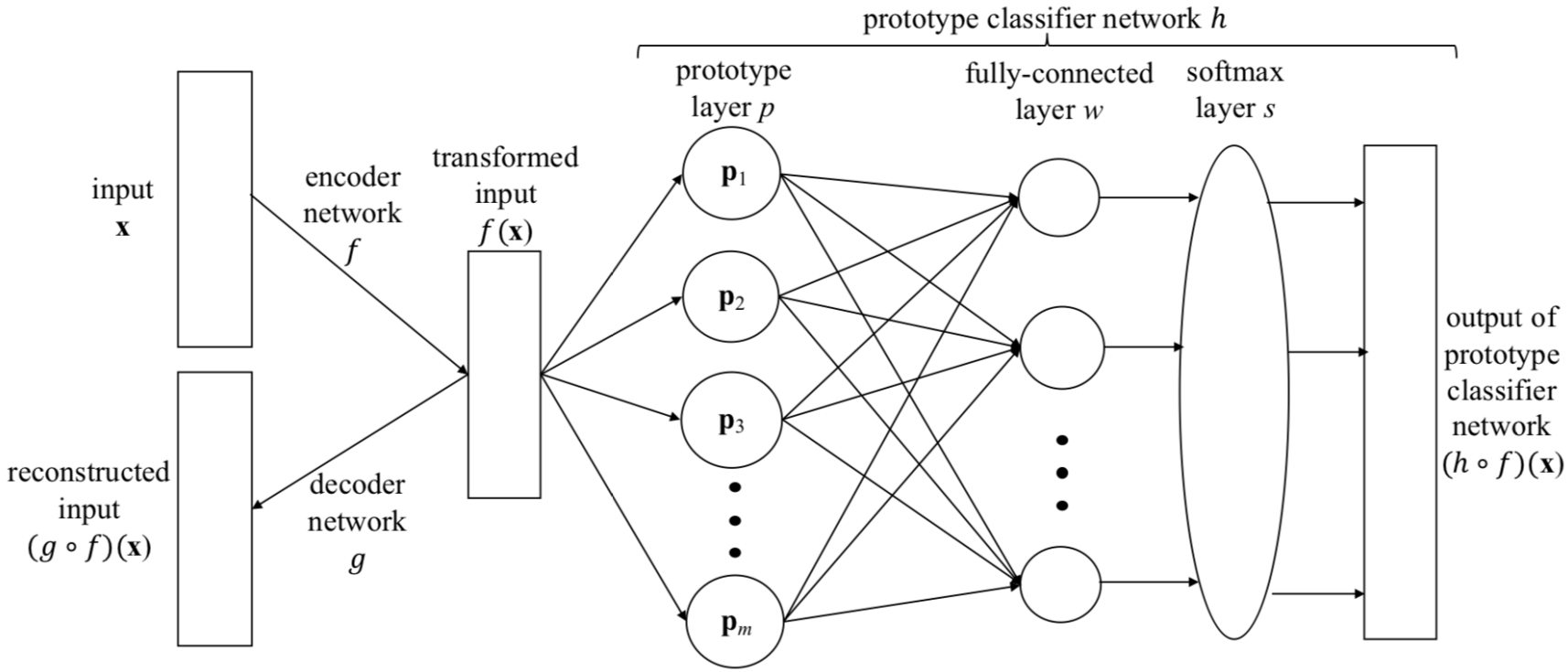


M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135?1144. ACM, 2016.

Prototype learning

- A prototype is a data instance that is representative of all the data
- Prototypes can improve the interpretability of complex data distributions (but usually they can not explain the data)
- Any clustering algorithm that returns actual data points as cluster centers would qualify for selecting prototypes
- Data points in areas with high data density are good prototypes

Prototype learning cont'd



Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin. *Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions*, AAAI, 2018

Prototype learning cont'd

Two interpretable regularization terms:

$$R_1(p_1, \ldots, p_m, D) = \frac{1}{m} \sum_{j=1}^{m} \min_{i \in [1,n]} \|p_j - f(x_i)\|_2^2, \tag{14}$$

$$R_2(p_1, \ldots, p_m, D) = \frac{1}{m} \sum_{j=1}^{m} \min_{j \in [1,n]} \|p_j - f(x_j)\|_2^2 \tag{15}$$

- Minimization of R_1 requires each prototype vector to be as close as possible to at least one of the training examples (in the latent space)
- Minimization of R_2 requires every (encoded) training example to be as close as possible to one of the prototypes

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin. *Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions*, AAAI, 2018

Sparse Models

$$\text{Loss function} + \lambda\text{Penalty} \tag{16}$$

- The L_2 penalty term is to avoid overfitting
- The L_1 penalty term induces sparsity
- Sparse (compact) models are considered to be more interpretable
- Penalty terms including the L_1: group penalties, hierarchical penalties, etc.

Rounding Methods

Input: X, Y
Output: weights associated with each (observed) value in X

function Project(w) $= \max(-R, \min(w, R))$
function RandomizedRounding(w,ϵ)
$b = \epsilon \left\lceil \frac{w}{\epsilon} \right\rceil$
$a = \epsilon \left\lfloor \frac{w}{\epsilon} \right\rfloor$
return b with probability $(w - a)/\epsilon$,
and a with probability $1 - (w - a)/\epsilon$

Organisation

Medical Data and Some Applications

State-of-the-art Supervised ML Methods

Interpretable Models

Challenges in Interpretable Supervised Learning

The Challenges

Constructing optimal logical models

- A model consisting of statements "or", "and", "if-then", etc.
- Often called *rule lists*
- Expert systems (1970's)
- Optimisation problem:

$$\min_{f \in \mathcal{F}} \left(\frac{1}{n} \mathbb{1}_{\{\text{training observation } i \text{ is misclassified by } f\}} + \lambda \times \text{size}(f) \right) \tag{17}$$

- The parameter λ is the classification error one would sacrifice in order to have one fewer term in the model; if λ is 0.01, it means we would sacrifice 1% training accuracy in order to reduce the size of the model by one.

from C. Rudin. *Please Stop Explaining Black Box Models for High-Stakes Decisions*, 2019

The Challenges cont'd

Construct optimal sparse scoring systems

$CHADS_2$ Score to assess stroke risk:

1. *Congestive Heart Failure*	1 point		⋯
2. *Hypertension*	1 point	+	⋯
3. *Age* $\geq$ *75*	1 point	+	⋯
4. *Diabetes Mellitus*	1 point	+	⋯
5. *Prior Stroke or Transient Ischemic Attack*	2 points	+	⋯
ADD POINTS FROM ROWS 1–5	**SCORE**	=	⋯

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

- Often used in medicine (and criminology)
- Optimisation problem

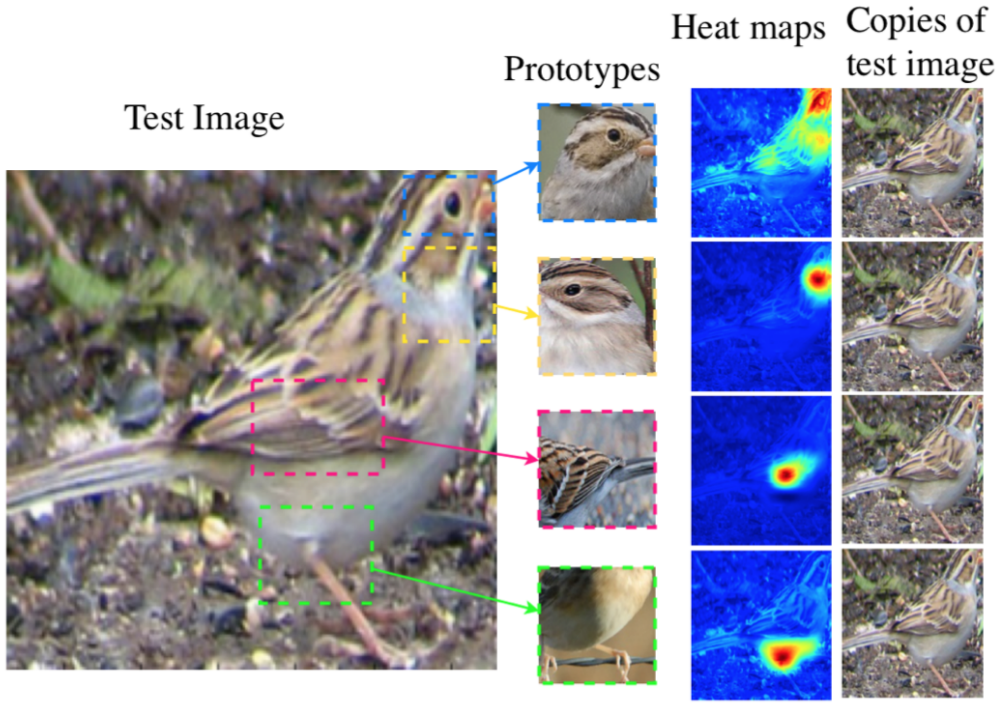
$$\min_{b_1,b_2,\ldots,b_p\in\{-10,-9,\ldots,9,10\}} \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp(-\sum_{j=1}^{p} b_j x_{ij})\right)+\lambda\sum_{j}\mathbb{1}_{\{b_j\neq 0\}} \tag{18}$$

- The model size is the number of non-zero coefficients, and λ is the trade-off parameter

from C. Rudin. *Please Stop Explaining Black Box Models for High-Stakes Decisions*. 2018

The Challenges cont'd

Define interpretability for specific domains and create methods accordingly including computer vision



From *Chen et al., 2018*: parts of the image are similar to prototypical parts of training examples.

from C. Rudin. *Please Stop Explaining Black Box Models for High-Stakes Decisions*, 2018