Interpretable Machine Learning: Scoring Systems, Causal Inference, and Network Reconstruction

Nataliya Sokolovska

NutriOmics Team, INSERM, Sorbonne University

Habilitation Defense
May, 13, 2019
Outline

Motivation

Interpretable Probabilistic Models

Probabilistic Causal Inference

Network Reconstruction

Conclusions and Perspectives
Motivation

Heterogeneous patient population

Heterogeneous data

Machine Learning Methods

Disease signature, disease stage, risk assessment, treatment decision, disease control, personalized medicine

Stratified subgroups
Problems and Solutions

1. How to predict a health status and stratify patients?
   - Interpretable models for prediction (of a disease)

2. How to prevent an event (e.g. an illness)?
   - Causal inference

3. Discover and explain a mechanism (of a pathology)
   - Network reconstruction

Aim: development of novel machine learning methods with theoretical guarantees
Motivation

**Interpretable Probabilistic Models**

**Probabilistic Causal Inference**

**Network Reconstruction**

**Conclusions and Perspectives**
Motivation: the DiaRem (Diabetes Prediction) Score

<table>
<thead>
<tr>
<th>Variable</th>
<th>Thresholds</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>&lt;40</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>40–49</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>50–59</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>&gt;60</td>
<td>3</td>
</tr>
<tr>
<td>Glycated hemoglobin</td>
<td>&lt;6.5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6.5–6.9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>7–8.9</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>&gt;9</td>
<td>6</td>
</tr>
<tr>
<td>Insuline</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>10</td>
</tr>
<tr>
<td>Other drugs</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>3</td>
</tr>
</tbody>
</table>

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

Real Medical Scores

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.
Scoring Systems: 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
Scores Learning

**Machine Learning point of view:**

- Problems are formulated and solved as linear integer tasks

- Bayesian optimisation is used to fit a model

- Linear methods (regressions) using gradient-based optimisation, with rounded coefficients
Automated Score Construction

1. Identification of related clinical variables

| age | glycated hemoglobin | insuline | other drugs |
|-----|---------------------|----------|-------------|-------------|
## Automated Score Construction

1. **Identification of related clinical variables**

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>40–49</td>
<td>50–59</td>
<td>&gt;60</td>
</tr>
</tbody>
</table>

2. **Meaningful thresholds for clinical variables**

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;6.5</td>
<td>6.5–6.9</td>
<td>7–8.9</td>
<td>&gt;9</td>
</tr>
</tbody>
</table>

Classify as **Remission** if sum of scores < 7

Classify as **Non-remission** if sum of scores ≥ 7
Automated Score Construction

1. Identification of related clinical variables

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>40–49</td>
<td>50 – 59</td>
<td>&gt;60</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

2. Meaningful thresholds for clinical variables

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;6.5</td>
<td>6.5 – 6.9</td>
<td>7 – 8.9</td>
<td>&gt;9</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;6.5</td>
<td>6.5 – 6.9</td>
<td>7 – 8.9</td>
<td>&gt;9</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

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

1. Identification of related clinical variables

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
</table>

2. Meaningful thresholds for clinical variables

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>&lt; 6.5</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>40–49</td>
<td>6.5 – 6.9</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>50 – 59</td>
<td>7 – 8.9</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>&gt;60</td>
<td>&gt; 9</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>age</th>
<th>glycated hemoglobin</th>
<th>insuline</th>
<th>other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>&lt;6.5</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>40–49</td>
<td>6.5 – 6.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50 – 59</td>
<td>7 – 8.9</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>&gt;60</td>
<td>&gt; 9</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

4. Find an optimal separator between two classes

Classify as Remission if sum of scores $<$ 7
Classify as Non-remission if sum of scores $\geq$ 7
Scoring Systems: Problem Formulation

- Training examples \( \{X_i, Y_i\}_{i=1}^N \)

- Discretized training examples \( \{Z_i, Y_i\}_{i=1}^N \), where \( Z \) is the interval (or one-hot) encoding of \( X \), and \( Y \)

- Score function is defined as \( \langle \theta, Z \rangle \), where \( \theta \) is a coefficient vector.

- A class can be predicted according to the conditional probability

\[
p(y = 1|Z) = \frac{1}{1 + \exp(-\langle \theta, Z \rangle)}.
\] (1)
Idea 1: Fused Lasso for Interpretable Models

We minimise the hinge loss

\[
\sum_{i=1}^{N} \ell(y_i, \theta \cdot z_i + b) + \lambda \sum_{j=1}^{\tilde{d}-1} |\theta_j - \theta_{j+1}|. \tag{2}
\]

As an optimisation (linear programming) problem:

\[
\min \left( \sum_{i=1}^{N} \xi_i + \sum_{j=1}^{\tilde{d}} \eta_j \right), \text{ such that } \tag{3}
\]

for all \(i\), \(y_i(\theta \cdot z_i + b) \geq 1 - \xi_i\), \(\tag{4}\)

for all \(j\), \(-\lambda \eta_j \leq \theta_j - \theta_{j+1} \leq \lambda \eta_j\), \(\tag{5}\)

\(\xi_i \geq 0, \theta_i \in \mathbb{N} \) for all \(i\). \(\tag{6}\)

Fused lasso shrinks similar variables to each other creating bins

Idea 2: Fully Corrective Binning

- **Binning as a feature selection task**
- **Add a new feature**, i.e., split one of the existing bins into two bins, if this operation minimizes the empirical risk:

  \[
  j, l, u, r = \arg\max_{\text{for all } j, l, u, r \in [l, u]} \left( \max\left( |(\nabla R)_{jlr}|, |(\nabla R)_{jru}| \right) \right),
  \]

  \[
  \theta = (\theta \cup \{\theta_{jlr}, \theta_{jru}\}) - \{\theta_{jl}\}.
  \]

- **Remove a feature**, i.e., merge two bins if does not degrade performance:

  \[
  j, l, u, q = \arg\min_{\text{for all } j, l, q, u, q \in [l, u]} \left( |\theta_{jlq} - \theta_{jqu}| \right),
  \]

  \[
  \theta = (\theta \cup \{\theta_{jlq}\}) - \{\theta_{jlq}, \theta_{jqu}\}.
  \]

The AdDiaRem

- New biomarkers (diabetes duration, number of drugs taken)

The distributions of the DiaRem and AdDiaRem scores

*J. Aron-Wisnewsky et al., Diabetologia, 2017*
Scoring Systems on Simulated Data
Interpretable Models: Ongoing and Future Work

- Learning under budget constraints (time, money, side effects)
- Cascade classifiers
- Discrete classifiers (randomized rounding)

---

M. Clertant, N. Sokolovska, Y. Chevaleyre, B. Hanczar. AISTATS, 2019
Interpretable Models: Ongoing and Future Work

- Learning under budget constraints (time, money, side effects)
- Cascade classifiers
- Discrete classifiers (randomized rounding)

Ongoing Work

- *Interpretable Cascade Classifiers with Abstention*
- Individualised dynamic protocols: a trade-off between the price (medical tests) and the prediction (diagnostic) accuracy

M. Clertant, N. Sokolovska, Y. Chevaleyre, B. Hanczar. AISTATS, 2019
Motivation

Interpretable Probabilistic Models

Probabilistic Causal Inference

Network Reconstruction

Conclusions and Perspectives
Motivation: Obesity and Causality

Does gastric bypass surgery prevent onset of diabetes?

- We already tried to *predict* remission/non-remission
- If our aim is to *prevent* (*and not to predict*) the type 2 diabetes: is the Gastric Bypass a good intervention?

from David Sontag
Causal Inference: a Machine Learning Viewpoint

- Machine learning influence
- Absence of time series

**Postulate** (For a bivariate case, $X$ and $Y$ variables):
If $X \rightarrow Y$, then the marginal distribution of the cause $P(X)$ and the conditional distribution of the effect given the cause $P(Y|X)$ are "independent" in the sense that $P(Y|X)$ contains no information about $P(X)$ and vice versa.

Interventions and Independent Mechanisms

\( A \) – altitude, and \( T \) – temperature

\[
p(a, t) = p(a|t)p(t) \quad \text{(factorization according to } T \rightarrow A) \\
p(t|a)p(a) \quad \text{(factorization according to } A \rightarrow T) 
\]

Consider the **effect of interventions**:

- ▶ Interventing on \( A \) has changed \( T \)
  - ▶ We climb higher
  - ▶ The temperature is lower

- ▶ Interventing on \( T \) has not changed \( A \)
  - ▶ We do not change the altitude
  - ▶ We build a massive heating system around the city that raises the temperature
Interventions and Independent Mechanisms

$A$ – altitude, and $T$ – temperature

\[ p(a, t) = \]
\[ p(a|t)p(t) \text{ (factorization according to } T \rightarrow A) \]
\[ p(t|a)p(a) \text{ (factorization according to } A \rightarrow T) \]
\[ p(t|a) \text{ and } p(a) \]

mechanism and probability of observations are independent

no influence of $p(a)$ on $p(t|a)$

Consider the effect of interventions:

- Intervening on $A$ has changed $T$
  - We climb higher
  - The temperature is lower

- Intervening on $T$ has not changed $A$
  - We do not change the altitude
  - We build a massive heating system around the city that raises the temperature
Measuring Independence

  \[D_{Y|X} = dCor(P(X), P(Y|X)); D_{X|Y} = dCor(P(Y), P(X|Y))\]

- CURE (Causal inference with Unsupervised inverse REgression, Sgouritsa et al., AISTATS, 2015)
  \[D_{X|Y} = \mathcal{L}_{X|Y}^{\text{unsup}} - \mathcal{L}_{X|Y}^{\text{sup}}; \quad D_{Y|X} = \mathcal{L}_{Y|X}^{\text{unsup}} - \mathcal{L}_{Y|X}^{\text{sup}}\]

  \[E[(X - \psi(Y))^2] \leq E[(Y - \phi(X))^2]\]
Measuring Independence

  
  \[ D_{Y|X} = dCor(P(X), P(Y|X)); D_{X|Y} = dCor(P(Y), P(X|Y)) \]

- CURE (Causal inference with Unsupervised inverse REgression, Sgouritsa et al., AISTATS, 2015)
  
  \[ D_{X|Y} = \mathcal{L}^{unsup}_{X|Y} - \mathcal{L}^{sup}_{X|Y}; \quad D_{Y|X} = \mathcal{L}^{unsup}_{Y|X} - \mathcal{L}^{sup}_{Y|X} \]

  
  \[ \mathbb{E}[(X - \psi(Y))^2] \leq \mathbb{E}[(Y - \phi(X))^2] \]

  
  \[ D_{X|Y} = \mathcal{L}^{semi-sup}_{X|Y} - \mathcal{L}^{sup}_{X|Y}; D_{Y|X} = \mathcal{L}^{semi-sup}_{Y|X} - \mathcal{L}^{sup}_{Y|X} \]
Causal Inference Algorithm given an Independence Measure

Input: Samples $X$ and $Y$, a threshold $\epsilon$
Output: Causality directions

STEP 1: Compute $P(X)$ and $P(Y|X)$ from data,
Estimate $D_{Y|X} = D(P(X), P(Y|X))$

STEP 2: Compute $P(Y)$ and $P(X|Y)$ from data,
Estimate $D_{X|Y} = D(P(Y), P(X|Y))$

STEP 3: Decide the edge direction:
\[
\text{if } D_{Y|X} - D_{X|Y} > \epsilon \text{ then} \\
\quad \text{Infer } X \rightarrow Y
\]

\[
\text{end if}
\]
\[
\text{if } D_{X|Y} - D_{Y|X} > \epsilon \text{ then} \\
\quad \text{Infer } Y \rightarrow X
\]

\[
\text{end if}
\]

Note that if $|D_{Y|X} - D_{X|Y}| < \epsilon$, then the approach can not provide any edge orientation.
Revealing Causal Relations between Groups of Variables

NutriOmics team, Pitié-Salpêtrière hospital examined 49 patients (10 groups of heterogeneous variables).

N. Sokolovska, K. Clément, J.-D. Zucker. Information Fusion, 2019
Future and Ongoing Work: Identifying Latent Variables and Common Hidden Causes

**Figure:** Simplified causal relationship between obesity and disease

(A) A direct causal effect of obesity on the occurrence of diseases. (B) Diet, physical activity, and other factors have a direct effect on obesity and diseases, but there is no direct effect of obesity on the risk of disease. In both cases, obesity is associated with a loss of disease-free years.
Motivation

Interpretable Probabilistic Models

Probabilistic Causal Inference

Network Reconstruction

Conclusions and Perspectives
A challenging application: the human gut microbiota

How to identify relevant meta-species interactions?

- Discover relations between genes, metagenomic species, metabolites, heterogeneous data, etc.
- High-dimensional setting, sparse and noisy data sets
- Biological systems: overlapping sub-units (metabolic or gene regulatory networks)
- Most of bacterial species have no reference genome
Huge Networks Reconstruction

▶ Constraint-based methods (PC)
  ▶ ascertain conditional independence from statistical tests; guaranteed to learn the Markov equivalent class; sensitive to noise
▶ Score-based methods (Hill-climbing)
  ▶ identify the model that fits the data best (maximization of a score); sensitive to noise
▶ Hybrid and mutual information-based approaches (3off2: S.Affeldt et al., 2016; Aracne: Margolin et al., 2006)
▶ Graphical Lasso (J. Friedman, T. Hastie, R. Tibshirani, 2014)
Consensus Network

- Learn a big network from **local reconstructions**
- **Spectral theory** → uncover structure in data (algebraic connectivity or Fiedler vector)
- **Eigenvector basis** provides a new representation that amplifies the similarity between related variables
- Each **eigenvector** individually incorporates variable membership to clusters (*Fiedler M., 1975; Newman MEJ, 2006*)
SCS (Spectral Consensus Strategy) relies on the elements of each eigenvector to identify path-related variables

- Reconstruct per eigenvector $v_k$
  - $m$ variables with highest positive $v_k$ elements
  - $m$ variables with highest negative $v_k$ elements
Spectral Consensus Strategy

SCS and human gut ecosystem reconstruction
(2,101 MGS, n=663)

Microbial co-presence ecosystem. Top 30% edges.
Ongoing and Future Work: Network Reconstruction

Applications in Systems biology

- Coordinator of AAP Défi Santé Numérique 2019 Modelling metabolism of intestinal microbiome by multi-“omics” statistical data integration
- Reconstruct metabolic networks from genome data
- Our aim: robust measures on directed graphs
- Graphical measures that reflect (correlate with) environmental conditions

Motivation

Interpretable Probabilistic Models

Probabilistic Causal Inference

Network Reconstruction

Conclusions and Perspectives
Some Conclusions and Perspectives

- (ANR JCJC *DiagnoLearn* 2018 – 2020)
  - Learning under budget
  - Discrete models
  - Heuristics for NP-hard optimisation problems

- Huge directed and undirected graphs
  - Covariate shift for causal inference
  - Spectral methods
  - Stable (feature selection) methods

- Real-world applications
  - Medicine and biology, metagenomic signatures for diseases
  - Materials science applications
  - Dynamic (temporal) data