Bayesian Networks for Interpretable Machine Learning and Optimization

Pedro Larrañaga

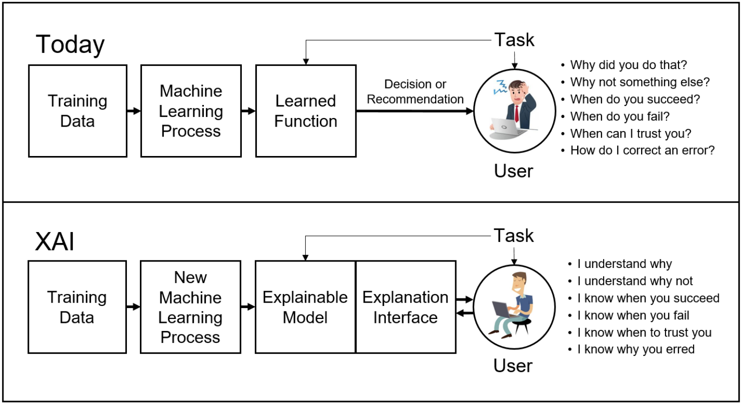
Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

3rd International Symposium on New Trend in Computational Intelligence, December 12, 2021

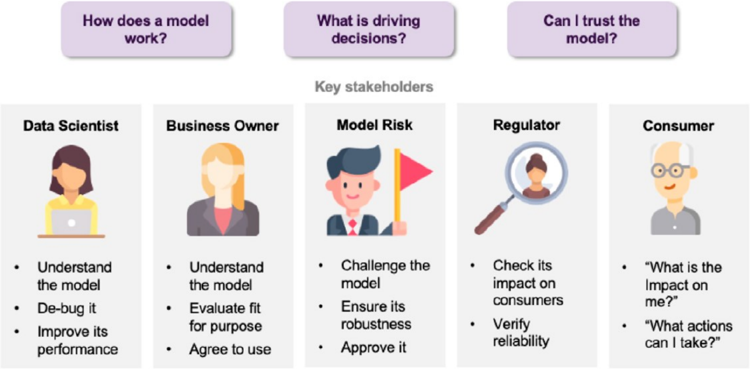
Outline

Outline

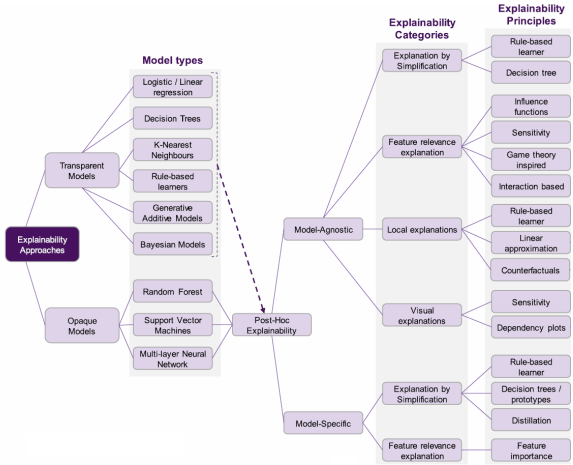
Explainable Artificial Intelligence (XAI) (Gunning, 2017)



Concerns faced by various stakeholders (Belle and Papantonis, 2021)



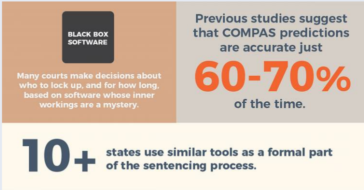
A taxonomy of XAI approaches (Belle and Papantonis, 2021)



- Rudin C (2019). Stop explaining black box models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206-215

Explaining black box models

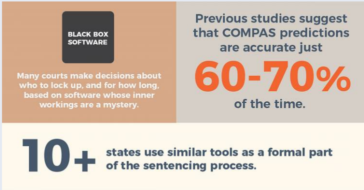
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (Northpointe, 2013)



- Secret formula for predicting criminal recidivism
- Unnecessarily complicated as it does not seem to be any more accurate than a very sparse decision tree

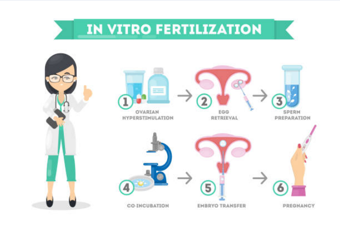
Explaining black box models

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (Northpointe, 2013)



- Secret formula for predicting criminal recidivism
- Unnecessarily complicated as it does not seem to be any more accurate than a very sparse decision tree

In vitro fertilization (Afnan et al. 2021)



- Black box models: Inability to perform shared decision-making with patients
- Accountability issues: Who is accountable when the model causes harm?

Interpretability (Lipton, 2016)

Human in the loop

- Interpretability stands for a human-level understanding of the inner working of the model
  - Simulatability refers to a model's ability to be simulated by a human. Simplicity alone is not enough /very large amount of simple rules versus a neural networks with no hidden layers). At the level of the entire model
  - Decomposability denotes the ability to break down a model into parts and then interpret these parts. At the level of individual components
  - Algorithmic transparency expresses the ability to understand the procedure the model goes through to generate its output. At the level of the training algorithm

Interpretability (Lipton, 2016)

Human in the loop

- Interpretability stands for a human-level understanding of the inner working of the model
  - Simulatability refers to a model's ability to be simulated by a human. Simplicity alone is not enough /very large amount of simple rules versus a neural networks with no hidden layers). At the level of the entire model
  - Decomposability denotes the ability to break down a model into parts and then interpret these parts. At the level of individual components
  - Algorithmic transparency expresses the ability to understand the procedure the model goes through to generate its output. At the level of the training algorithm

Interpret to

- Justify the decisions of the intelligent system to other people
- Understand its weakness
- Discover new knowledge
- Robustness. Are minor perturbations (or the presence of missing or noisy data) susceptible to change the outcome of the intelligent system?
- Bias. Can we detect biases in the data that unfairly penalize groups of individuals?
- Improvement. How can the prediction model be improved?
- Transferability. Under which circumstances the prediction model for one application domain can be applied (transferred) to another application domain?
- Human comprehensibility. Are we able to explain the model's algorithmic machinery to an expert? And to a non-expert?

Introduction

References

- Afnan MAM, et al. (2021). Ethical implementation of artificial intelligenc to select embryos in in vitro fertilization. *Proceedings of the Fourth AAAI/ACM Conference on Artificial Intelligence*
- Belle V, Papantonis I (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 1, article 688969
- European Commision (2020). *White Paper on Artificial Intelligence: An European Approach to Excellence and Trust*. Brussels
- Davies A et al. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, 600, 7074
- Gunning D (2017). *Explainable Artificial Intelligence*. DARPA/I20 Program
- High-Level Expert Group on AI (2019). *Ethics Guidelines for Trustworthy AI*. Brussels
- Lipton ZC (2016). The mythos of model interpretability. *Communications of the ACM*, 61 (10)
- Northpointe (2013). *Practitioner's Guide to COMPAS Core*. Technical Report, Northpointe
- Rudin C (2019). Stop explaining black box models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206-215
- Wachter S, Mittelstadt B, Floridi L (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76-99

Outline

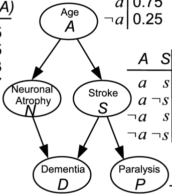
Bayesian networks

DAG + CPTs

- Conditional independence: **W** and **T** are conditionally independent given **Z** $\Leftrightarrow p(\mathbf{W}|\mathbf{T},\mathbf{Z}) = p(\mathbf{W}|\mathbf{Z})$
- Directed acyclic graph (DAG)
- Conditional probability tables (CPTs)
- $p(X_1,\ldots,X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{Pa}(X_i))$

A	p(A)
a	0.75
$\neg a$	0.25

A	N	p(N|A)
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97



A	S	p(S|A)
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	p(D|N,S)
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	p(P|S)
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

$p(A,N,S,D,P) = p(A)p(N|A)p(S|A)p(D|N,S)p(P|S)$

Bayesian networks

DAG + CPTs

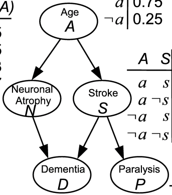
- Conditional independence: **W** and **T** are conditionally independent given **Z** $\Leftrightarrow p(\mathbf{W}|\mathbf{T},\mathbf{Z}) = p(\mathbf{W}|\mathbf{Z})$
- Directed acyclic graph (DAG)
- Conditional probability tables (CPTs)
- $p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{Pa}(X_i))$

Inference

- Exact: variable elimination, message passing
- Approximate: sequential simulation and MCMC

A	$p(A)$
a	0.75
$\neg a$	0.25

A	N	$p(N\|A)$
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97

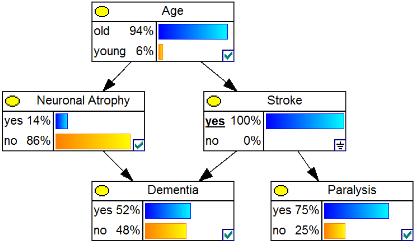


A	S	$p(S\|A)$
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	$p(D\|N,S)$
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	$p(P\|S)$
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

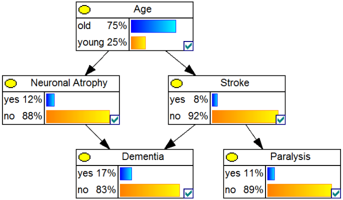
$p(A, N, S, D, P) = p(A)p(N|A)p(S|A)p(D|N,S)p(P|S)$



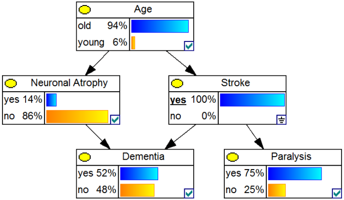
$p(X_i|\texttt{Stroke=yes})$

Bielza, Larrañaga, 2020

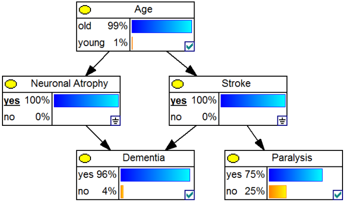
Conditional independence. An example



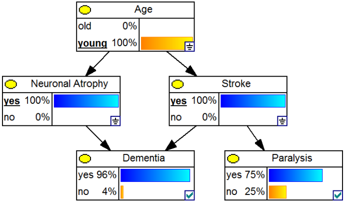
$p(X_i)$



$p(X_i|\texttt{Stroke=yes})$



$p(X_i|\texttt{Stroke=yes, Neural Atropy=yes})$

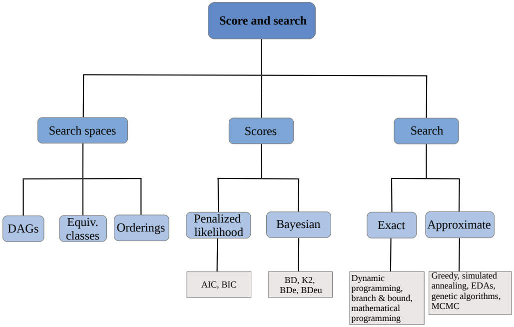


$p(X_i|\texttt{Stroke=yes, Neural Atropy=yes, Age=young})$

Learning Bayesian networks from data

Two elements

- Parameters $p(X_i = x_i \mid \mathbf{Pa}(X_i) = \mathbf{pa}_i^j)$: MLE or Bayesian
- Structure: conditional independence tests or by optimizing a *score*



Scores

- Penalized likelihood: avoid structural overfitting
- Bayesian: $\arg\max_{\mathcal{G}} p(\mathcal{G}|\mathcal{D})$, with

$$p(\mathcal{G}|\mathcal{D}) \propto \overbrace{p(\mathcal{D}|\mathcal{G})}^{\text{marginal likeli.}} \ \overbrace{p(\mathcal{G})}^{\text{prior}}, \text{ with}$$

$$p(\mathcal{D}|\mathcal{G}) = \int \underbrace{p(\mathcal{D}|\mathcal{G}, \boldsymbol{\theta})}_{\text{likelihood}} \ \underbrace{f(\boldsymbol{\theta}|\mathcal{G})}_{\text{prior param.}} \ d\boldsymbol{\theta}$$

Bayesian networks for machine learning

MACHINE LEARNING

Multi-output regression
- Problem transformation methods
  - Single-target method
  - Multi-target regressor stacking
  - Regressor chains
- Algorithm adaptation methods
  - Multi-output suppor vector regression
  - Kernel methods
  - Multi-target regression trees
  - Rule method
  - Gaussian Bayesian networks

Clustering
- Non-probabilistic
  - Hierarchical
    - Agglomerative
    - Divisive
  - Partitional
    - K-means
    - K-medians
    - K-modes
    - Fuzzy C-means
    - Self-organizing map
    - Spectral clustering
    - K-medoids
    - Affinity propagation
    - K-plane clustering
    - Fuzzy C-shell
    - DBSCAN
- Probabilistic
  - Finite-mixture models
  - Bayesian networks

Multiview-clustering
- Agglomerative clustering
- Density-based
- Principal component analysis
  - Canonical correlation analysis
- Spectral clustering
- Co-regularization
- Ensemble clustering
- Bayesian networks

Feature subset selection
- Filter approaches
  - Univariate filters
  - Multivariate filters
- Wrapper methods
  - Bayesian networks-based EDAs
- Embedded methods
- Hybrid methods

Discovery associations
- Association rules
- Bayesian networks

Anomaly detection
- Probabilistic approaches
- Distance-based
- Reconstruction-based
- Domain-based
- Information theory-based
- Bayesian network likelihood

Multi-dimensional classifiers
- Problem transformation methods
  - Binary relevance
  - Classifier chains
  - RAkEL
  - LPBR
  - Label powerset
- Algorithm adaptation methods
- Multi-dimensional Bayesian network classifiers

Supervised classification
- Non-probabilistic
  - Nearest neighbors
  - Classification trees
  - Rule induction
  - Artificial neural networks
  - Support vector machines
- Probabilistic
  - Discriminant analysis
  - Logistic regression
  - Bayesian network classifiers
- Metaclassifiers
  - Fusion of outputs
  - Stacked generalization
  - Cascading
  - Bagging
  - Random forest
  - Boosting
  - Hybrid classifiers

Bayesian networks for machine learning

MACHINE LEARNING

Multi-output regression
- Problem transformation methods
  - Single-target method
  - Multi-target regressor stacking
  - Regressor chains
- Algorithm adaptation methods
  - Multi-output suppor vector regression
  - Kernel methods
  - Multi-target regression trees
  - Rule method
  - **Gaussian Bayesian networks**

Clustering
- Non-probabilistic
  - Hierarchical
    - Agglomerative
    - Divisive
  - Partitional
    - K-means
    - K-medians
    - K-modes
    - Fuzzy C-means
    - Self-organizing map
    - Spectral clustering
    - K-medoids
    - Affinity propagation
    - K-plane clustering
    - Fuzzy C-shell
    - DBSCAN
- Probabilistic
  - Finite-mixture models
  - **Bayesian networks**

Multiview-clustering
- Agglomerative clustering
- Density-based
- Principal component analysis
  - Canonical correlation analysis
- Spectral clustering
- Co-regularization
- Ensemble clustering
- **Bayesian networks**

Feature subset selection
- Filter approaches
  - Univariate filters
  - Multivariate filters
- Wrapper methods
  - **Bayesian networks-based EDAs**
- Embedded methods
- Hybrid methods

Supervised classification
- Non-probabilistic
  - Nearest neighbors
  - Classification trees
  - Rule induction
  - Artificial neural networks
  - Support vector machines
- Probabilistic
  - Discriminant analysis
  - Logistic regression
  - **Bayesian network classifiers**
- Metaclassifiers
  - Fusion of outputs
  - Stacked generalization
  - Cascading
  - Bagging
  - Random forest
  - Boosting
  - Hybrid classifiers

Discovery associations
- Association rules
- **Bayesian networks**

Anomaly detection
- Probabilistic approaches
- Distance-based
- Reconstruction-based
- Domain-based
- Information theory-based
- **Bayesian network likelihood**

Multi-dimensional classifiers
- Problem transformation methods
  - Binary relevance
  - Classifier chains
  - RAkEL
  - LPBR
  - Label powerset
- Algorithm adaptation methods
- **Multi-dimensional Bayesian network classifiers**

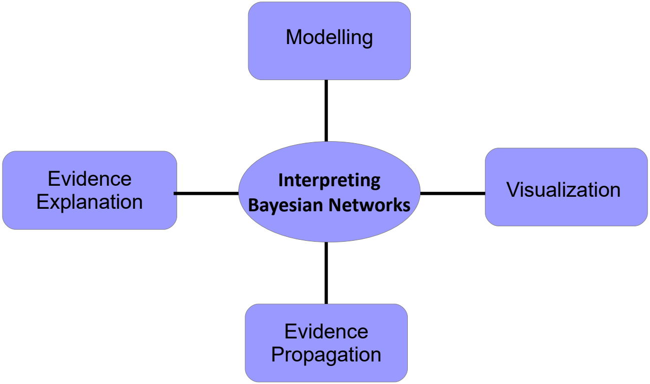
Bayesian networks

References

- Bielza C, Larrañaga P (2020). *Data-Driven Computational Neuroscience. Machine Learning and Statistical Models*. Cambridge University Press
- Castillo E, Gutierrez JM, Hadi A (1997). *Expert Systems and Probabilistic Network Models*. Springer
- Darwiche A (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press
- Jensen F, Nielsen TD (2007). *Bayesian Networks and Decision Graphs*. Springer
- Koller D, Friedman N (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press
- Lauritzen S (1996). *Graphical Models*. Oxford University Press
- Maathuis M, Drton M, Lauritzen S, Wainwright M (2019). *Handbook of Graphical Models*. CRC Press
- Neapolitan (2003). *Learning Bayesian Networks*. Prentice Hall
- Pearl J (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann
- Sucar E (2015). *Probabilistic Graphical Models: Principles and Applications*. Springer

Outline

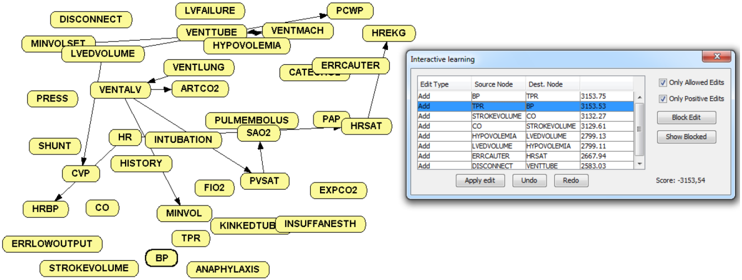
Interpreting Bayesian networks



Interactive learning of Bayesian networks (Bermejo et al. 2012)

OpenMarkov

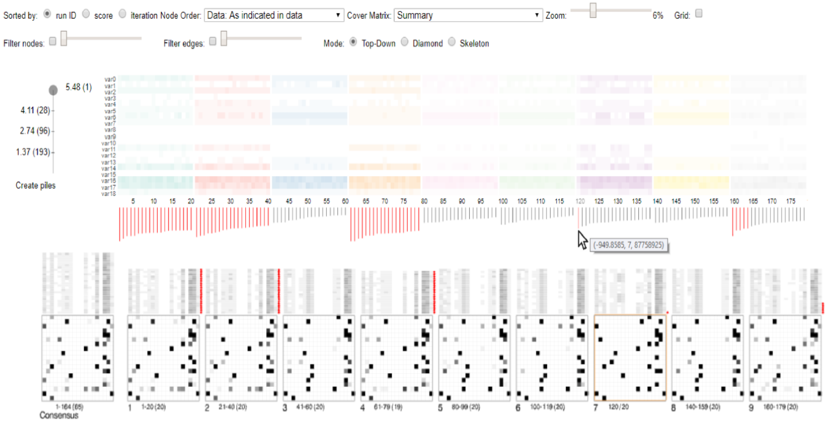
- The option to run the algorithms in a step-by-step fashion
- Interactive learning is performed by having two windows: one showing the DAG and another one showing the proposed edits
- The user can select any edit from the list, not necessarily the one having the highest score, and the change will be immediately displayed on the network window. Alternatively, the user can add or remove any link from the DAG. In both cases, the scores will be recalculated and a new list will be proposed
- It is focussed on score + search approaches to structure learning with a greedy strategy



Consensus of Bayesian network structures (Kennedy et al. 2018)

BayesPiles

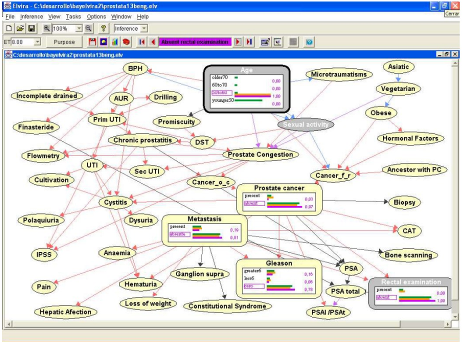
- Software for exploring, combining and comparing large collections of Bayesian networks learnt during the search
- Heuristics for the search: greedy search and simulated annealing
- Interactive consensus process: human in the loop



Visualization (Lacave and Diez, 2000; Elvira Consortium, 2002)

Elvira software: `http://www.ia.uned.es/~elvira/instalar/Elvira.zip`

- Importance factor of a node: number between 0 and 10 given by the expert
- The nodes whose importance factor is greater than, or equal to, the expansion threshold are automatically expanded
- Automatic colouring the arcs of the DAG, in order to offer qualitative insight about the CPTs
- Verbal explanation: "`The disease [Name] has the following RISK FACTORS [List of risk factors]. It presents with the following SIGNS [List of signs] and SYMPTOMS [List of symptoms]. There are several TESTS to confirm or discard its diagnosis [List of tests]`"

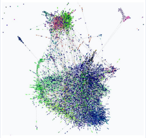


Visualization (Michiels et al. 2020)

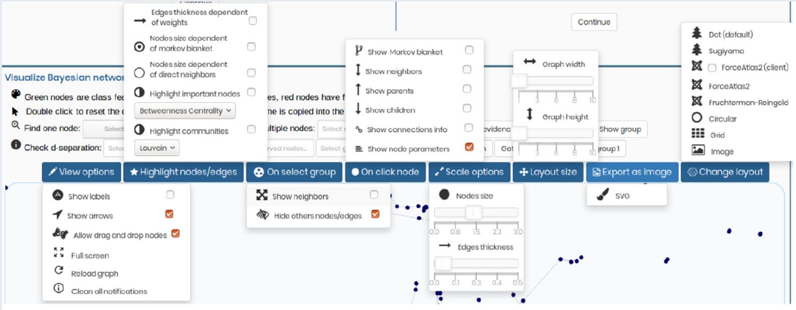
BayeSuites `https://neurosuites.com/`

- A web framework for learning, visualizing, and interpreting Bayesian networks that scale to tens of thousands of nodes
- For a Bayesian networks with 20,000 nodes and 20,000 arcs:
  - Time to model 10-15 s
  - Time to layout $<$ 60 s

Louvain algorithm



Visualization options

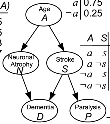


Evidence propagation. Exact methods (Bielza and Larrañaga, 2020)

The “Dementia” BN

A	$p(A)$
a	0.75
$\neg a$	0.25

A	N	$p(N\|A)$
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97



A	S	$p(S\|A)$
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	$p(D\|N,S)$
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	$p(P\|S)$
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

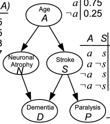
$p(A, N, S, D, P) =$
$p(A)p(N|A)p(S|A)p(D|N, S)p(P|S)$

Evidence propagation. Exact methods (Bielza and Larrañaga, 2020)

The "Dementia" BN

A	p(A)
a	0.75
$\neg a$	0.25

A	N	p(N|A)
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97



A	S	p(S|A)
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	p(D|N,S)
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	p(P|S)
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

$p(A, N, S, D, P) =$
$p(A)p(N|A)p(S|A)p(D|N,S)p(P|S)$

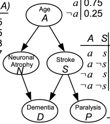
Brute force for $p(D)$

$$\begin{aligned}
p(D) &= \sum_{A,N,S,P} p(A,N,S,P,D) \\
&= \sum_{A,N,S,P} p(A)p(N|A)p(S|A)p(D|N,S)p(P|S) \\
&= \sum_{A} p(A) \sum_{N} p(N|A) \sum_{S} p(S|A)p(D|N,S) \sum_{P} p(P|S)
\end{aligned}$$

128 multiplications and 16 additions are required to yield $p(d)$

Evidence propagation. Exact methods (Bielza and Larrañaga, 2020)

The “Dementia” BN



A	p(A)
a	0.75
$\neg a$	0.25

A	N	p(N|A)
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97

A	S	p(S|A)
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	p(D|N,S)
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	p(P|S)
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

$p(A,N,S,D,P) = p(A)p(N|A)p(S|A)p(D|N,S)p(P|S)$

Brute force for $p(D)$

$$
\begin{aligned}
p(D) &= \sum_{A,N,S,P} p(A,N,S,P,D) \\
&= \sum_{A,N,S,P} p(A)p(N|A)p(S|A)p(D|N,S)p(P|S) \\
&= \sum_{A} p(A) \sum_{N} p(N|A) \sum_{S} p(S|A)p(D|N,S) \sum_{P} p(P|S)
\end{aligned}
$$

128 multiplications and 16 additions are required to yield $p(d)$

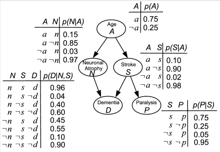
Variable elimination (Zhang and Poole, 1994) for $p(S|\neg d)$

Consider $\mathcal{L} = \{f_A(A), f_N(N,A), f_S(S,A), f_P(P,S), f_D(\neg d, S, N)\}$ and the ordering P-A-N

$$p(S|\neg d) \propto \sum_{N} p(\neg d|N,S) \sum_{A} p(N|A)p(S|A)p(A) \sum_{P} p(P|S)$$

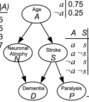
Evidence propagation. Exact methods (Bielza and Larrañaga, 2020)

The "Dementia" BN



Evidence propagation. Exact methods (Bielza and Larrañaga, 2020)

The "Dementia" BN



A	p(A)
a	0.75
$\neg a$	0.25

A	N	p(N|A)
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97

A	S	p(S|A)
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	p(D|N,S)
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

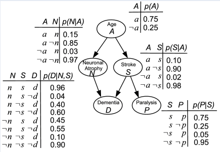
S	P	p(P|S)
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

Message passing algorithm (Lauritzen and Spiegelhalter, 1988)

- Moralize the Bayesian network
- Triangulate the moral graph and output the cliques (nodes of the junction tree)
- Create the junction tree and assign initial potentials to each clique

Evidence propagation. Exact methods (Bielza and Larrañaga, 2020)

The "Dementia" BN



A	p(A)
a	0.75
$\neg a$	0.25

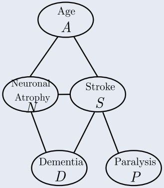
A	N	p(N|A)
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97

A	S	p(S|A)
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	p(D|N,S)
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	p(P|S)
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

Moral graph

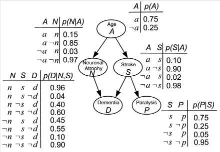


Message passing algorithm (Lauritzen and Spiegelhalter, 1988)

- Moralize the Bayesian network
- Triangulate the moral graph and output the cliques (nodes of the junction tree)
- Create the junction tree and assign initial potentials to each clique

Evidence propagation. Exact methods (Bielza and Larrañaga, 2020)

The "Dementia" BN



A	p(A)
a	0.75
$\neg a$	0.25

A	N	p(N|A)
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97

A	S	p(S|A)
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

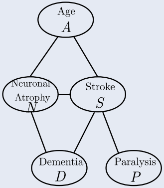
N	S	D	p(D|N,S)
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	p(P|S)
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

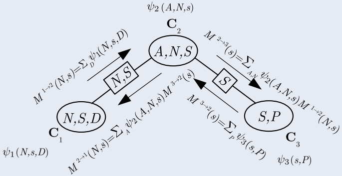
Message passing algorithm (Lauritzen and Spiegelhalter, 1988)

- Moralize the Bayesian network
- Triangulate the moral graph and output the cliques (nodes of the junction tree)
- Create the junction tree and assign initial potentials to each clique

Moral graph



Junction tree and the message passing



Evidence explanation

$\mathbf{X} = (X_1, ..., X_n)$ random variables in the Bayesian network; $\mathbf{E} \subset \mathbf{X}$ evidence; $\mathbf{U} = \mathbf{X} \setminus \mathbf{E}$ unobserved variables; $\mathbf{H} \subset \mathbf{U}$ variables of interest; $\mathbf{U} = \mathbf{H} \cup \mathbf{I}$, C class variable

Types of queries

- Posterior probability of a target variable $X_i \subset \mathbf{U}$ given the evidence $\mathbf{e} \rightarrow p(x_i|\mathbf{e})$
- Posterior joint of a set of target variables $\mathbf{H} \subset \mathbf{U}$ given the evidence $\mathbf{e} \rightarrow p(\mathbf{h}|\mathbf{e})$
- Abductive reasoning: most likely configuration event that best explains the evidence (Kwisthout 2011)
  - Total abduction: most probable explanation (MPE), the search for all the unobserved variables $\rightarrow$ $\mathbf{u}^* = \arg\max_{\mathbf{u}} p(\mathbf{u}|\mathbf{e})$
  - Partial abduction: maximum a posteriori (MAP), that is search for a subset of unobserved variables $\rightarrow$ $\mathbf{h}^* = \arg\max_{\mathbf{h}} p(\mathbf{h}|\mathbf{e})$
  - k most likely explanations: k MPE and k MAP
- Most relevant explanation (MRE) (Yuan et al. 2011): assignment of a subset of the unobserved variables that maximizes its generalized Bayes factor $\rightarrow \mathbf{h}^* = \arg\max_{\mathbf{h}} \frac{p(\mathbf{e}|\mathbf{h})}{p(\mathbf{e}|\bar{\mathbf{h}})}$
- MAP-independence explanation (Kwisthout 2021) $\rightarrow \mathbf{h}^* = \arg\max_{\mathbf{h}} p(\mathbf{h}|\mathbf{e}) = \arg\max_{\mathbf{h}} \sum_{i \in \Omega(I)} p(\mathbf{H} = \mathbf{h}, \mathbf{I} = \mathbf{i}|\mathbf{e})$. The goal is to partition the set $\mathbf{I}$ into variables $\mathbf{I}^+$ that are relevant to establishing the best explanation, and variables $\mathbf{I}^-$ that are irrelevant
- Counterfactual reasoning in classification problems (Albini et al. 2020). Given $\mathbf{x}$ such that $p(C = +|\mathbf{x}) > p(C = -|\mathbf{x})$, the goal is to find $\mathbf{x}'$ very similar to $\mathbf{x}$, such that $p(C = +|\mathbf{x}') < p(C = -|\mathbf{x}')$

References (i)

Modelling. Visualization. Evidence Propagation. Evidence Explanation

- Albini E, Rago A, Baroni P, Toni F (2020). Relation-based counterfactual explanations for Bayesian network classifiers. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 451-457
- Bermejo I, Oliva J, Díez FJ, Arias M (2012). Interactive learning of Bayesian networks using `OpenMarkov`. *Sixth European Workshop on Probabilistic Graphical Models*, 27-34
- Bielza C, Larrañaga P (2020). *Data-Driven Computational Neuroscience. Machine Learning and Statistical Models*. Cambridge University Press
- Elvira Consortium (2002). Elvira: An enviroment for probabilistic graphical models. *Proceedings of the First European Wokshop on Probabilistic Graphical Models*, 222-230
- Kennedy J, Archambault D, Bach B, Smith VA (2018). BayesPiles: Visualization support for Bayesian network structure learning. *ACM Transactions on Intelligent Systems and Technology*, 10(1), 1-5
- Kwisthout J (2011). Most probable explanations in Bayesian networks: Complexity and tractability. *International Journal of Approximate Reasoning* 52(9), 1452-1469
- Kwisthout J (2021). Explainable AI using MAP-independence. *Lecture Notes in Computer Science* 12897, 243-254

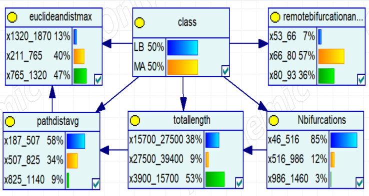
References (ii)

Modelling. Visualization. Evidence Propagation. Evidence Explanation

- Lacave C, Diez FJ (2000). A review of explanation methods for Bayesian networks. *Knowledge Engineering Review*, 17(2),107-127
- Lauritzen S, Spiegelhalter D (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2), 157-224
- Michiels M, Larrañaga P, Bielza C (2020). BayeSuites: An open web framework for massive Bayesian networks focused on neuroscience. *Neurocomputing*, 428, 166-181
- Yuan C, Lim H, Lu T-C (2011). Most relevant explanation in Bayesian networks. *Journal of Machine Learning Research*, 42, 309-352
- Zhang N, Poole D (1994). A simple approach to Bayesian network computations, *Proceedings of the 10th Biennial Canadian Conference on Artificial Intelligence*, 171-178

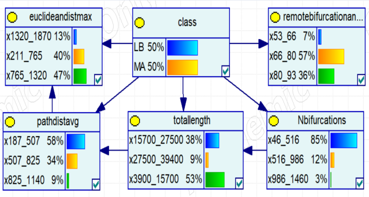
Morphological classification of interneurons (Mihaljević et al. 2015)

Marginal probabilities

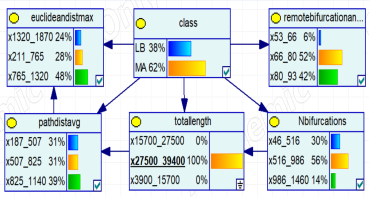


Morphological classification of interneurons (Mihaljević et al. 2015)

Marginal probabilities

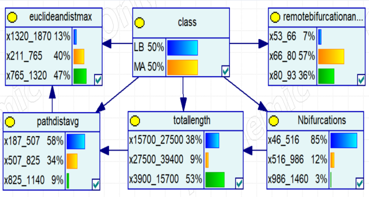


Evidence in one predictor variable

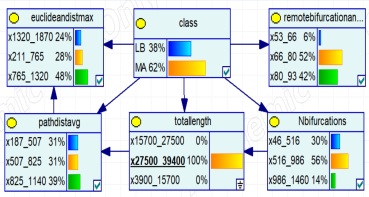


Morphological classification of interneurons (Mihaljević et al. 2015)

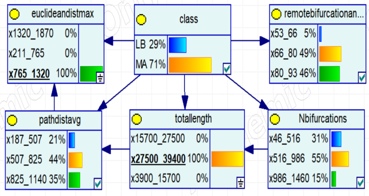
Marginal probabilities



Evidence in one predictor variable

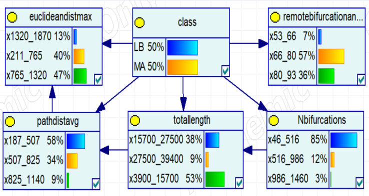


Evidence in two predictor variables

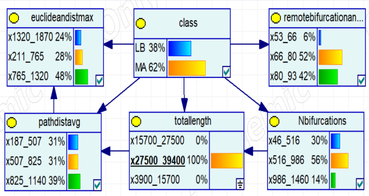


Morphological classification of interneurons (Mihaljević et al. 2015)

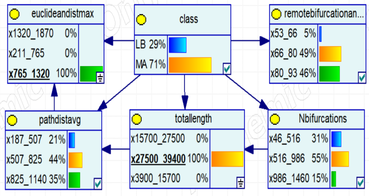
Marginal probabilities



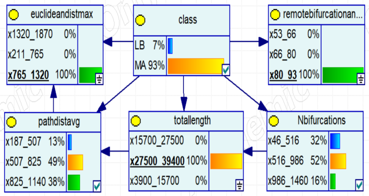
Evidence in one predictor variable



Evidence in two predictor variables

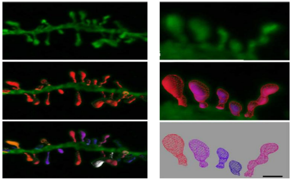


Evidence in three predictor variables



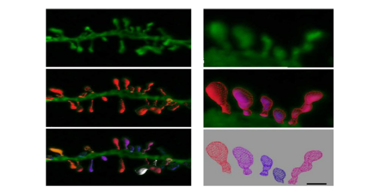
Clustering of dendritic spines (Luengo-Sanchez et al. 2018)

Human dendritic spines

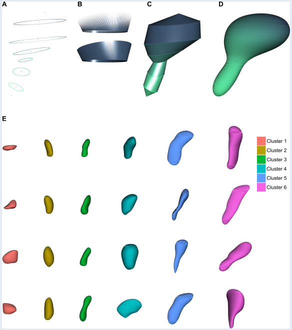


Clustering of dendritic spines (Luengo-Sanchez et al. 2018)

Human dendritic spines

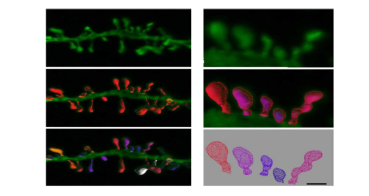


Virtual spines simulated from the model

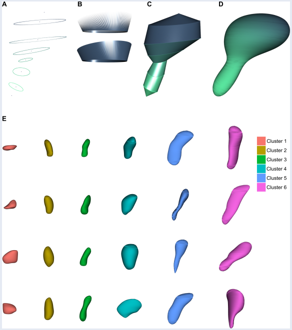


Clustering of dendritic spines (Luengo-Sanchez et al. 2018)

Human dendritic spines



Virtual spines simulated from the model

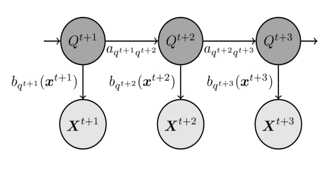


Mixture of Gaussian Bayesian networks

- In a multivariate Gaussian mixture model: $f(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x};\boldsymbol{\mu}_k)$ each mixture density is given by: $f_k(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\}$
- In a mixture of Gaussian Bayesian networks each component is expressed as a Gaussian Bayesian network

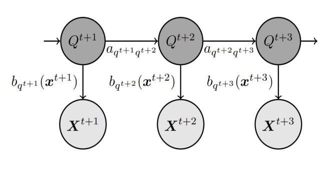
Autoregressive asymmetric linear Gaussian hidden Markov models (AR-AsLG-HMM) (Puerto-Santana et al. 2021)

An HMM as a Bayesian network

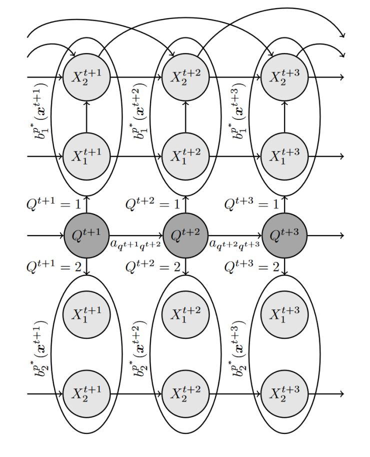


Autoregressive asymmetric linear Gaussian hidden Markov models (AR-AsLG-HMM) (Puerto-Santana et al. 2021)

An HMM as a Bayesian network

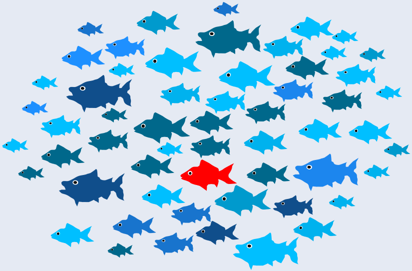


Graphical representation of an AR-AsLG-HMM model



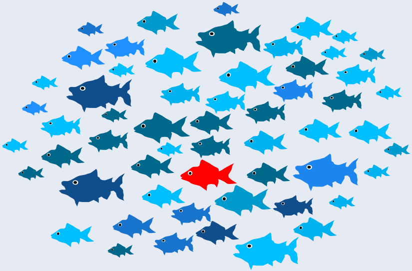
Anomaly detection

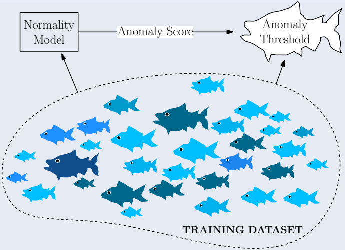
Anomaly detection via likelihood of new instances (Larrañaga et al. 2018)



Anomaly detection

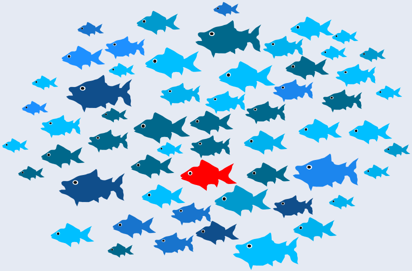
Anomaly detection via likelihood of new instances (Larrañaga et al. 2018)

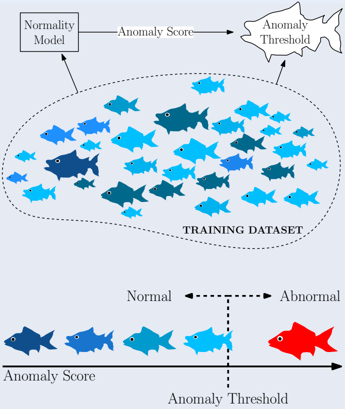




Anomaly detection

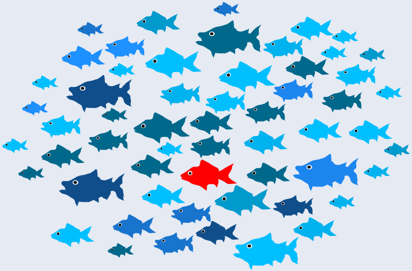
Anomaly detection via likelihood of new instances (Larrañaga et al. 2018)

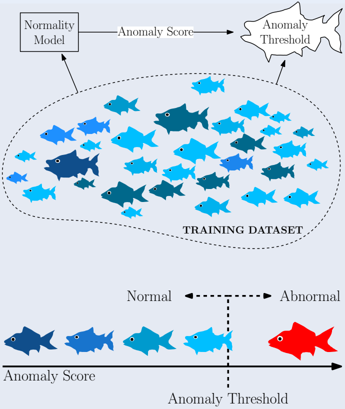




Anomaly detection

Anomaly detection via likelihood of new instances (Larrañaga et al. 2018)





1. Compute a probabilistic model based on (dynamic) Bayesian networks for the normal instances
2. Establish a threshold in this joint probability distribution
3. Compare the likelihood of the new instance with the likelihood threshold

Incorporating previous knowledge

Physics informed Bayesian networks

Two options:

1. Incorporate the expert knowledge into the structure of the Bayesian network
2. Simulation from a non-linear process defined by a system of ordinary differential equations + Learning the structure of the Bayesian network from this dataset (Quesada et al. 2021)

Bayesian approaches

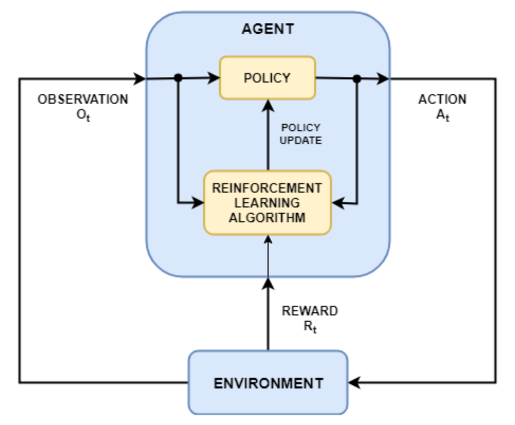
- Conjugate prior distributions over the parameters. Dirichlet for mulltinomial. Normal-Wishart for Gaussian
- Prior distribution over the structures

Transfer learning (Velázquez et al. 2008)

- Transferring parameters with probability aggregation methods combining probabilities estimated from the target domain with those obatined from the auxiliary data
- Transferring structures by means of conditional independence tests using a weighted sum of conditional independence measures

Reinforcement learning

Interpreting reinforcement learning policies with Bayesian networks



- Modeling agent learning experience with Bayesian networks (Jin et al. 2011)
- Causal reinforcement learning (Zhang and Bareinboim, 2020)

Machine learning tasks

References (i)

- Bielza C, Li G, Larrañaga P (2011). Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, 52, 705-727
- Borchani H, Bielza C, Martínez-Martín P, Larrañaga P (2012). Multidimensional Bayesian network classifiers applied to predict the European quality of life-5 dimensions (EQ-5D) from the 39-item Parkinson's disease questionnaire (PDQ-39), *Journal of Biomedical Informatics*, 45, 1175-1184
- Jin Z, Jin J, Song J (2011). Learning from experience: A Bayesian network based reinforcement learning approach. *International Conference on Information Computing and Applications*, 407-414
- Larrañaga P, Atienza D, Diaz-Rojo J, Puerto-Santana CE, Ogbechie A, Bielza C (2018). *Industrial Applications of Machine Learning*. CRC Press
- Luengo-Sanchez S, Fernaud-Espinosa I, Bielza C, Benavides-Piccione R, Larrañaga P, J. DeFelipe (2018). 3D morphology-based clustering and simulation of human pyramidal cell dendritic spines. *PLOS Computational Biology*, 14(6), e1006221
- Luengo-Sanchez S, Larrañaga P, Bielza C (2019). A directional-linear Bayesian network and its application for clustering and simulation of neural somas. *IEEE Access*, 7(1), 69907-69921
- Mihaljević B, Benavides-Piccione R, Bielza C, DeFelipe J, Larrañaga P (2015). Bayesian network classifiers for categorizing cortical GABAergic interneurons. *Neuroinformatics*, 13(2), 192208

Machine learning tasks

References (ii)

- Puerto-Santana C, Larrañaga P, Bielza C (2021) Autoregressive asymmetric linear Gaussian hidden Markov models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10.1109/TPAMI.2021.3068799
- Quesada D, Larrañaga P, Bielza C, Font P (2021). Piece-wise forecasting of non-linear time series with dynamic Bayesian networks. In preparation
- Varando G, Bielza C, Larrañaga P (2015). Decision boundary for discrete Bayesian network classifiers. *Journal of Machine Learning Research*, 16, 2725-2749
- Velásquez R, Sucar LE, Morales EF (2008). Transfer learning for Bayesian networks. *Lecture Notes in Artificial Intelligence* 5290, 93102
- Zhang J, Bareinboim E (2020). Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. *Proceedings of the 37th International Conference on Machine Learning*, 11012-11022

Outline

Optimization

Heuristic search strategies

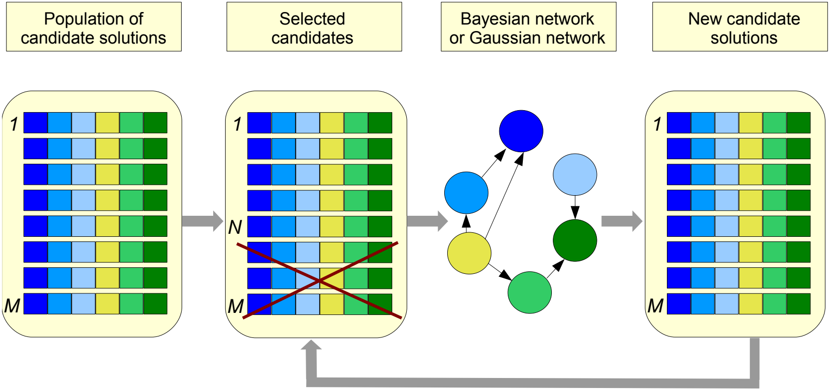
Deterministic heuristics

- Sequential feature selection
- Sequential forward feature selection
- Sequential backward elimination
- Greedy hill climbing
- Best first
- Plus-L-Minus-r algorithm
- Floating search selection
- Tabu search
- Branch and bound

Non-deterministic heuristics

- Single-solution metaheuristics:
  - Simulated annealing
  - Las Vegas algorithm
  - Greedy randomized adaptive search procedure
  - Variable neighborhood search
- Population-based metaheuristics:
  - Scatter search
  - Ant colony optimization
  - Particle swarm optimization
  - Evolutionary algorithms:
    - Genetic algorithms
    - **Estimation of distribution algorithms**
    - Differential evolution
    - Genetic programming
    - Evolution strategies

Estimation of distribution algorithms (Larrañaga and Lozano 2002)



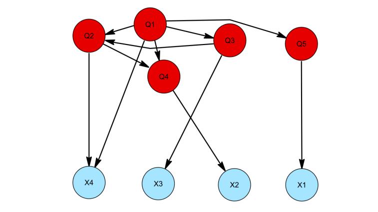
The Rashomon effect

- Rashomon effect (Breiman et al. 2001): A storytelling and writing method in cinema meant to provide different perspectives
- Rashomon set (Fisher et al. 2019; Dong and Rudin, 2020): A reduced set of individuals in the last generation

Estimation of distribution algorithms

Multi-objective estimation of distribution algorithms (Karshenas et al. 2014)

Joint modeling of objectives and variables for the 5-objective WFG1 optimization problem



A multi-objective optimization problem

$$\min_{\boldsymbol{x}} \quad \boldsymbol{Q}(\boldsymbol{x}) = (Q_1(\boldsymbol{x}), \ldots, Q_m(\boldsymbol{x}))$$

$$\text{subject to} \quad \begin{cases} \boldsymbol{x} \in \mathcal{D} \subseteq \mathbb{R}^n \\ \boldsymbol{Q} \in \mathcal{Q} \subseteq \mathbb{R}^m \end{cases}$$

The WFG1 multi-objective optimization problem

$$\begin{cases} Q_1(\boldsymbol{x}) = a + 2 \cdot h_1(g_2(x_1), g_2(x_2), g_2(x_3)) \\ Q_2(\boldsymbol{x}) = a + 4 \cdot h_2(g_2(x_1), g_2(x_2), g_2(x_3)) \\ Q_3(\boldsymbol{x}) = a + 6 \cdot h_3(g_2(x_1), g_2(x_2), g_2(x_3)) \\ Q_4(\boldsymbol{x}) = a + 8 \cdot h_4(g_2(x_1), g_2(x_2), g_2(x_3)) \\ Q_5(\boldsymbol{x}) = a + 10 \cdot h_5(g_2(x_1)) \\ a = g_1(x_5, \ldots, x_{16}) \end{cases}$$

Estimation of distribution algorithms

References

- Breiman L (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199-231
- Dong J, Rudin C (2020). Exploring the cloud of variable importance for the set of aal good models. *Nature Machine Intelligence*, 2(12), 810-824
- Fisher A, Rudin C, Dominici F (2019), All models are wrong, but many are useful: Learning a variable's importance by studing an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1-81
- Karshenas H, Santana R, Bielza C, Larrañaga P (2014). Multi-objective estimation of distribution algorithm based on joint modeling of objectives and variables. *IEEE Transactions on Evolutionary Computation*, 18 (4), 519-542
- Larrañaga P, Lozano JA (2002) *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Kluwer Academic Publishers

Outline

Conclusions

- Explainable AI is not enough for high stakes decisions
- Interpretable AI (simulatability, decomposability, algorithmic transparency) necessary
- Bayesian networks as a framework providing interpretability for machine learning and optimization

Further topics

- Interpreting other probabilistic graphical models
  - Sum-product networks
  - Influence diagrams
  - Probabilistic generative adversarial networks
  - Markov networks
  - Conditional random fields
- Interpreting Bayesian networks for temporal data
  - Dynamic Bayesian networks
  - Temporal Bayesian networks
  - Continuous time Bayesian networks
- Interpreting causal Bayesian networks

BAYESIAN NETWORKS FOR INTERPRETABLE MACHINE LEARNING AND OPTIMIZATION

Pedro Larrañaga

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

3rd International Symposium on New Trend in Computational Intelligence, December 12, 2021