Probabilistic Graphical Model Structure Learning: Application to Multi-Label Classification

PhD defense

Maxime Gasse

Supervised by: Alex Aussem and Haytham Elghazel

LIRIS DM2L, UMR 5205 CNRS
Université Lyon 1

13 January 2017

Thesis context

EU funding: ENIAC Joint Undertaking

- Integrated Solutions for Agile Manufacturing in High-mix Semiconductor Fabs
- 28 european partners leaded by STMicroelectronics

Thesis context

EU funding: ENIAC Joint Undertaking

- Integrated Solutions for Agile Manufacturing in High-mix Semiconductor Fabs
- 28 european partners leaded by STMicroelectronics

Bayesian network structure learning for process control in the semi-conductor industry.

Thesis context

EU funding: ENIAC Joint Undertaking

- Integrated Solutions for Agile Manufacturing in High-mix Semiconductor Fabs
- 28 european partners leaded by STMicroelectronics

Bayesian network structure learning for process control in the semi-conductor industry.

Research contributions in:

- BN structure learning (ECML 2012, ESWA 2014, IWBBIO 2014)
- Multi-label classification (ICML 2015, ECML 2016)
- Irreducible label factors (PGM 2016)

Outline

Probabilistic Graphical Models

What is a PGM?

Graphical: represents a set of independence constraints.

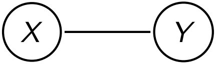
X and Y independent

X and Y dependent

What is a PGM?

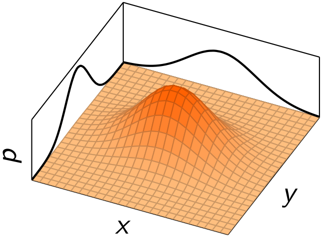
Graphical: represents a set of independence constraints.

X and Y independent

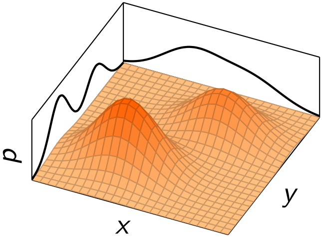


X and Y dependent

Probabilistic: encodes a probability distribution.



$p(x, y) = p(x)p(y)$



$p(x, y) \neq p(x)p(y)$

What is a PGM?

Independence model

Conditional independence relations:

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}).$$

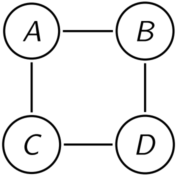
What is a PGM?

Independence model

Conditional independence relations:

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}).$$

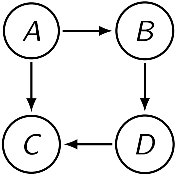
Undirected



$A \perp\!\!\!\perp D \mid \{B, C\}$

$B \perp\!\!\!\perp C \mid \{A, D\}$

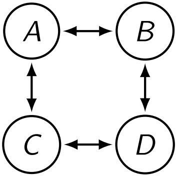
Directed



$A \perp\!\!\!\perp D \mid B$

$B \perp\!\!\!\perp C \mid \{A, D\}$

Bidirected



$A \perp\!\!\!\perp D \mid \emptyset$

$B \perp\!\!\!\perp C \mid \emptyset$

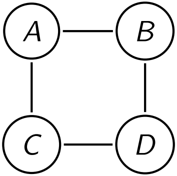
What is a PGM?

Independence model

Conditional independence relations:

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff p(\mathbf{x}, \mathbf{y} | \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) p(\mathbf{y}|\mathbf{z}).$$

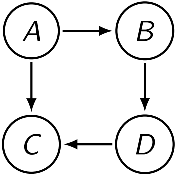
Undirected



$A \perp\!\!\!\perp D \mid \{B, C\}$

$B \perp\!\!\!\perp C \mid \{A, D\}$

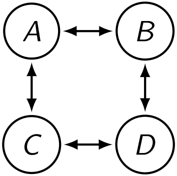
Directed



$A \perp\!\!\!\perp D \mid B$

$B \perp\!\!\!\perp C \mid \{A, D\}$

Bidirected



$A \perp\!\!\!\perp D \mid \emptyset$

$B \perp\!\!\!\perp C \mid \emptyset$

Different expressive powers.

What is a PGM?

A large family

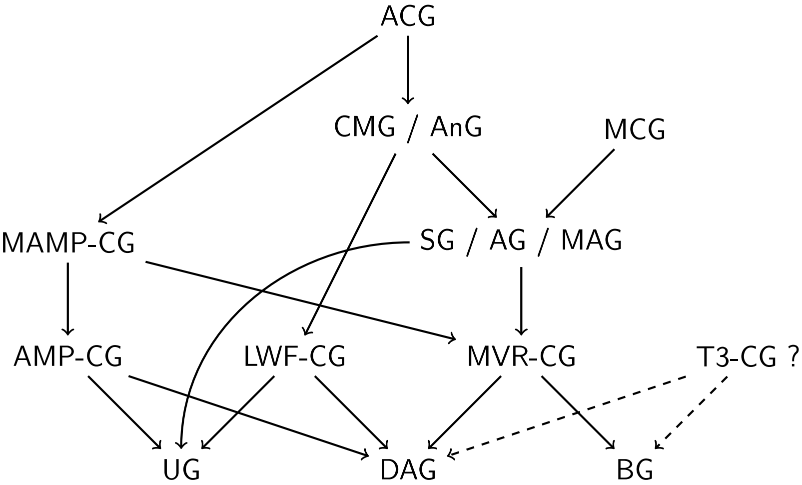


Figure: PGMs by order of inclusion (in terms of expressive power).

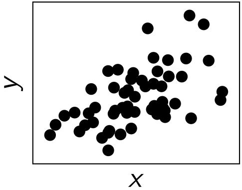
What is structure learning?

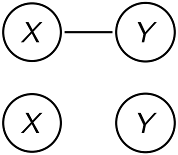
Learn a graph from a data set.

[1] D. M. Chickering, D. Heckerman, and C. Meek (2004). Large-Sample Learning of Bayesian Networks is NP-Hard.

What is structure learning?

Learn a graph from a data set.





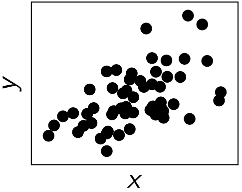
or

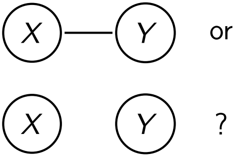
?

[1] D. M. Chickering, D. Heckerman, and C. Meek (2004). Large-Sample Learning of Bayesian Networks is NP-Hard.

What is structure learning?

Learn a graph from a data set.



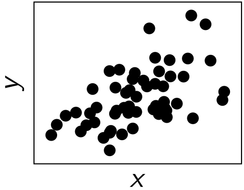


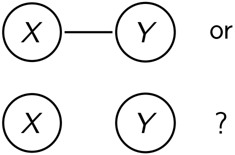
Why structure learning?

[1] D. M. Chickering, D. Heckerman, and C. Meek (2004). Large-Sample Learning of Bayesian Networks is NP-Hard.

What is structure learning?

Learn a graph from a data set.





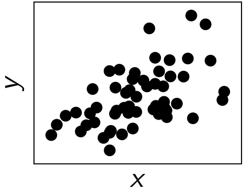
Why structure learning?

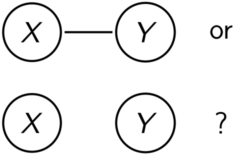
- model selection: sparse/dense graph = simple/complex model;

[1] D. M. Chickering, D. Heckerman, and C. Meek (2004). Large-Sample Learning of Bayesian Networks is NP-Hard.

What is structure learning?

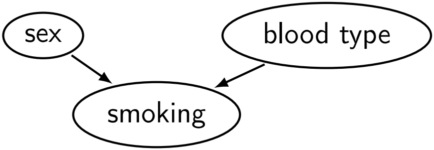
Learn a graph from a data set.





Why structure learning?

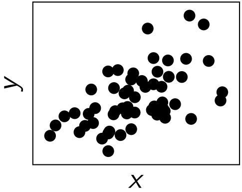
- model selection: sparse/dense graph = simple/complex model;
- interpretation:

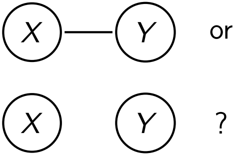


[1] D. M. Chickering, D. Heckerman, and C. Meek (2004). Large-Sample Learning of Bayesian Networks is NP-Hard.

What is structure learning?

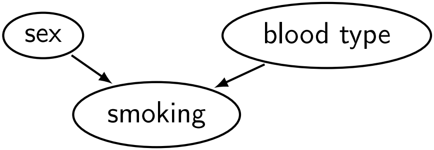
Learn a graph from a data set.





Why structure learning?

- model selection: sparse/dense graph = simple/complex model;
- interpretation:



NP-hard in general[1].

[1] D. M. Chickering, D. Heckerman, and C. Meek (2004). Large-Sample Learning of Bayesian Networks is NP-Hard.

What is structure learning?
Constraint-based approach

~~Score-based~~ / constraint-based:

What is structure learning?

Constraint-based approach

~~Score-based~~ / constraint-based:

- extract constraints: $A \perp\!\!\!\perp C \mid B$, $A \not\perp\!\!\!\perp C \mid \emptyset \ldots$;
- build a graph that respects these constraints.

What is structure learning?

Constraint-based approach

~~Score-based~~ / constraint-based:

- extract constraints: $A \perp\!\!\!\perp C \mid B$, $A \not\perp\!\!\!\perp C \mid \emptyset \ldots$;
- build a graph that respects these constraints.

Statistical tests, e.g. mutual information

$$I(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \propto \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{\mathbf{z} \in \mathcal{Z}} n_{\mathbf{x},\mathbf{y},\mathbf{z}} \log \frac{n_{\mathbf{x},\mathbf{y},\mathbf{z}} n_{\mathbf{z}}}{n_{\mathbf{x},\mathbf{z}} n_{\mathbf{y},\mathbf{z}}}.$$

What is structure learning?

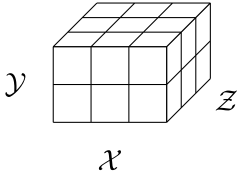
Constraint-based approach

~~Score-based~~ / constraint-based:

- extract constraints: $A \perp\!\!\!\perp C \mid B$, $A \not\perp\!\!\!\perp C \mid \emptyset \ldots$;
- build a graph that respects these constraints.

Statistical tests, e.g. mutual information

$$I(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \propto \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{\mathbf{z} \in \mathcal{Z}} n_{\mathbf{x},\mathbf{y},\mathbf{z}} \log \frac{n_{\mathbf{x},\mathbf{y},\mathbf{z}} n_{\mathbf{z}}}{n_{\mathbf{x},\mathbf{z}} n_{\mathbf{y},\mathbf{z}}}.$$



Keep **X**, **Y** and **Z** as small as possible!

What is structure learning?

Constraint-based approach

Do we need to perform all tests ?

[2] A. P. Dawid (1979). Conditional Independence in Statistical Theory.

What is structure learning?

Constraint-based approach

Do we need to perform all tests ?

Conditional independence properties = deductive system

semi-graphoid[2] (any p)

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$	Symmetry
$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$	Decomposition
$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}$	Weak Union
$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$	Contraction

[2] A. P. Dawid (1979). Conditional Independence in Statistical Theory.

What is structure learning?

Constraint-based approach

Do we need to perform all tests ?

Conditional independence properties = deductive system

semi-graphoid[2] (any p)

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$ — Symmetry

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ — Decomposition

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}$ — Weak Union

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$ — Contraction

graphoid ($p > 0$)

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$ — Intersection

[2] A. P. Dawid (1979). Conditional Independence in Statistical Theory.

What is structure learning?

Constraint-based approach

Do we need to perform all tests ?

Conditional independence properties = deductive system

semi-graphoid[2] (any p)

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$ — Symmetry

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ — Decomposition

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}$ — Weak Union

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$ — Contraction

graphoid ($p > 0$)

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$ — Intersection

compositional graphoid

$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}$ — Composition

---

[2] A. P. Dawid (1979). Conditional Independence in Statistical Theory.

Multi-Label Classification

What is MLC?

To which categories (plural) does an image belong?

What is MLC?

To which categories (plural) does an image belong?

(desert, mountains, sea, sunset, trees)

What is MLC?

To which categories (plural) does an image belong?

00110

10000

00001

00001

01100

10001

00011

00100

(desert, mountains, sea, sunset, trees)

What is MLC?

Probabilistic framework

Binary multi-output supervised learning: $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \{0, 1\}^m$,

$$\mathbf{h} : \mathcal{X} \rightarrow \mathcal{Y}.$$

What is MLC?

Probabilistic framework

Binary multi-output supervised learning: $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \{0, 1\}^m$,

$$\mathbf{h} : \mathcal{X} \to \mathcal{Y}.$$

Bayes-optimal prediction for $\mathbf{x} \iff$ minimal expected loss

$$\mathbf{h}^\star(\mathbf{x}) = \arg\min_{\hat{\mathbf{y}}} \sum_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) \times L(\hat{\mathbf{y}}, \mathbf{y}).$$

What is MLC?

Probabilistic framework

Binary multi-output supervised learning: $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \{0,1\}^m$,

$$\mathbf{h} : \mathcal{X} \to \mathcal{Y}.$$

Bayes-optimal prediction for $\mathbf{x} \iff$ minimal expected loss

$$\mathbf{h}^\star(\mathbf{x}) = \arg\min_{\hat{\mathbf{y}}} \sum_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) \times L(\hat{\mathbf{y}}, \mathbf{y}).$$

Very challenging:

- learn $p(\mathbf{y} \mid \mathbf{x}) \implies O(2^m)$ parameters;
- obtain $\mathbf{h}^\star(\mathbf{x}) \implies O(4^m)$ computations.

What is MLC?

Loss functions

$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, how far are $\hat{\mathbf{y}}$ and $\mathbf{y}$?

- Hamming loss $= \frac{1}{m} \sum_{i=1}^{m} [\hat{y}_i \neq y_i]$
- Zero-one loss $= [\hat{\mathbf{y}} \neq \mathbf{y}]$
- F-loss $= 1 - 2 \times \hat{\mathbf{y}} \cdot \mathbf{y} / (\hat{\mathbf{y}} \cdot \hat{\mathbf{y}} + \mathbf{y} \cdot \mathbf{y})$

[3] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier (2011). An Exact Algorithm for F-Measure Maximization.

What is MLC?

Loss functions

$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, how far are $\hat{\mathbf{y}}$ and $\mathbf{y}$?

- Hamming loss $= \frac{1}{m} \sum_{i=1}^{m} [\hat{y}_i \neq y_i]$
- Zero-one loss $= [\hat{\mathbf{y}} \neq \mathbf{y}]$
- F-loss $= 1 - 2 \times \hat{\mathbf{y}} \cdot \mathbf{y} / (\hat{\mathbf{y}} \cdot \hat{\mathbf{y}} + \mathbf{y} \cdot \mathbf{y})$

Affects MLC complexity:

	parameters		inference	
L_H	$p(y_i \| \mathbf{x})$	$O(m)$	$\arg\max_{\mathbf{y}} \prod_{i=1}^{m} p(y_i \| \mathbf{x})$	$O(m)$
$L_{0/1}$	$p(\mathbf{y} \| \mathbf{x})$	$O(2^m)$	$\arg\max_{\mathbf{y}} p(\mathbf{y} \| \mathbf{x})$	$O(4^m)$
L_F	$p(y_i \times \mathbf{y} \cdot \mathbf{y} \| \mathbf{x})$	$O(m^2)$	GFM[3]	$O(m^3)$

[3] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier (2011). An Exact Algorithm for F-Measure Maximization.

What is MLC?

Loss functions

$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, how far are $\hat{\mathbf{y}}$ and $\mathbf{y}$?

- Hamming loss $= \frac{1}{m} \sum_{i=1}^{m} [\hat{y}_i \neq y_i]$
- Zero-one loss $= [\hat{\mathbf{y}} \neq \mathbf{y}]$
- F-loss $= 1 - 2 \times \hat{\mathbf{y}} \cdot \mathbf{y} / (\hat{\mathbf{y}} \cdot \hat{\mathbf{y}} + \mathbf{y} \cdot \mathbf{y})$

Affects MLC complexity:

	parameters		inference	
L_H	$p(y_i \mid \mathbf{x})$	$O(m)$	$\arg\max_{\mathbf{y}} \prod_{i=1}^{m} p(y_i \mid \mathbf{x})$	$O(m)$
$L_{0/1}$	$p(\mathbf{y} \mid \mathbf{x})$	$O(2^m)$	$\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$	$O(4^m)$
L_F	$p(y_i \times \mathbf{y} \cdot \mathbf{y} \mid \mathbf{x})$	$O(m^2)$	GFM[3]	$O(m^3)$

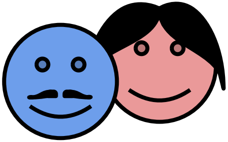
$\Longrightarrow$ PGMs particularly useful under $L_{0/1}$.

[3] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier (2011). An Exact Algorithm for F-Measure Maximization.

What is MLC?

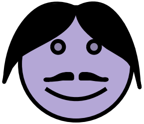
Loss functions

A quick example: who is in the picture?



Alice and Bob.

a	b	$p(a, b\|\mathbf{x})$	expected loss L_H	$L_{0/1}$
0	0	.02	.87	.99
0	1	.11	.49	.88
1	0	.12	.50	.89
1	1	.76	**.12**	**.24**



Alice or Bob?

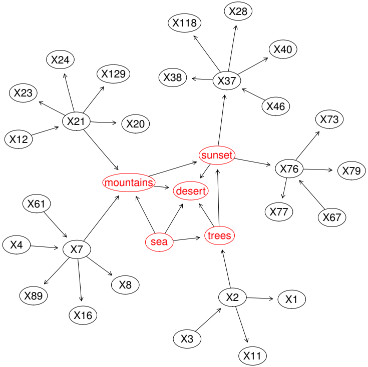
a	b	$p(a, b\|\mathbf{x})$	expected loss L_H	$L_{0/1}$
0	0	.02	.53	.98
0	1	.46	.49	**.54**
1	0	.44	.51	.56
1	1	.08	**.47**	.92

Why using PGMs?

Graphical structure $\iff$ constraints on $p(\mathbf{y}|\mathbf{x})$

Why using PGMs?

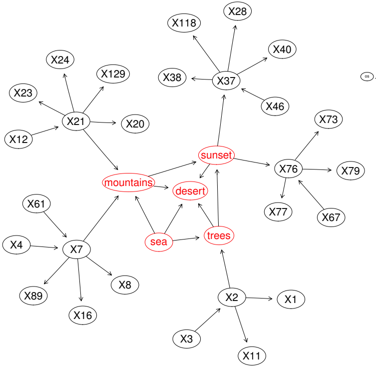
Graphical structure $\Longleftrightarrow$ constraints on $p(\mathbf{y}|\mathbf{x})$



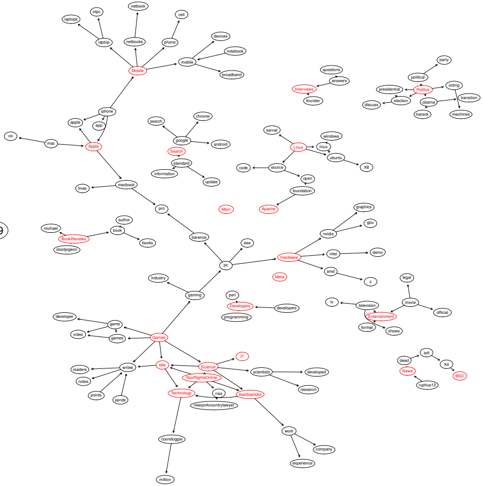
image

Why using PGMs?

Graphical structure $\Longleftrightarrow$ constraints on $p(\mathbf{y}|\mathbf{x})$



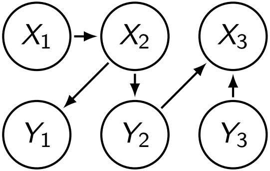
image



slashdot

Why using PGMs?

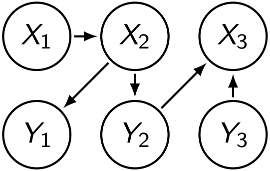
Disjoint factorization



$$p(\mathbf{y}|\mathbf{x}) = p(y_1|\mathbf{x}) \times p(y_2, y_3|\mathbf{x})$$

Why using PGMs?

Disjoint factorization



$$p(\mathbf{y}|\mathbf{x}) = p(y_1|\mathbf{x}) \times p(y_2, y_3|\mathbf{x})$$

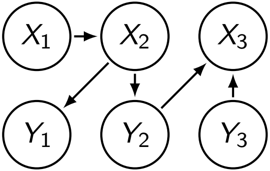
MLC under $L_{0/1}$:

$$\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg_{\mathbf{y}} \left[\max_{y_1} p(y_1|\mathbf{x}) \times \max_{y_2, y_3} p(y_2, y_3|\mathbf{x}) \right]$$

$$O(2^3) \implies O(2^1 + 2^2)$$

Why using PGMs?

Disjoint factorization



$$p(\mathbf{y}|\mathbf{x}) = p(y_1|\mathbf{x}) \times p(y_2, y_3|\mathbf{x})$$

MLC under $L_{0/1}$:

$$\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg_{\mathbf{y}} \left[\max_{y_1} p(y_1|\mathbf{x}) \times \max_{y_2, y_3} p(y_2, y_3|\mathbf{x}) \right]$$

$$O(2^3) \implies O(2^1 + 2^2)$$

Simplifies parameter learning and inference.

Why using PGMs?

Disjoint factorization

We want an irreducible disjoint factorization of $p(\mathbf{y}|\mathbf{x})$.

Definition

A label factor (LF) is a subset $\mathbf{Y}_F \subseteq \mathbf{Y}$ such that $\mathbf{Y}_F \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_F \mid \mathbf{X}$.
An irreducible label factor (ILF) is non-empty and contains no other non-empty LF.

[4] M. Gasse, A. Aussem, and H. Elghazel (2014). A hybrid algorithm for Bayesian network structure learning with application to multi-label learning.

[5] C. Bielza, G. Li, and P. Larrañaga (2011). Multi-dimensional classification with Bayesian networks.

[6] M. Gasse and A. Aussem (2016). Identifying the irreducible disjoint factors of a multivariate probability distribution.

Why using PGMs?

Disjoint factorization

We want an irreducible disjoint factorization of $p(\mathbf{y}|\mathbf{x})$.

Definition

A label factor (LF) is a subset $\mathbf{Y}_F \subseteq \mathbf{Y}$ such that $\mathbf{Y}_F \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_F \mid \mathbf{X}$.
An irreducible label factor (ILF) is non-empty and contains no other non-empty LF.

Initial idea: extract ILFs from a BN structure[4][5].

---

[4] M. Gasse, A. Aussem, and H. Elghazel (2014). A hybrid algorithm for Bayesian network structure learning with application to multi-label learning.

[5] C. Bielza, G. Li, and P. Larrañaga (2011). Multi-dimensional classification with Bayesian networks.

[6] M. Gasse and A. Aussem (2016). Identifying the irreducible disjoint factors of a multivariate probability distribution.

Why using PGMs?

Disjoint factorization

We want an irreducible disjoint factorization of $p(\mathbf{y}|\mathbf{x})$.

Definition

A label factor (LF) is a subset $\mathbf{Y}_F \subseteq \mathbf{Y}$ such that $\mathbf{Y}_F \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_F \mid \mathbf{X}$. An irreducible label factor (ILF) is non-empty and contains no other non-empty LF.

Initial idea: extract ILFs from a BN structure[4][5].

BN structure learning is hard, can we just learn ILFs?

- Yes, much simpler[6].

---

[4] M. Gasse, A. Aussem, and H. Elghazel (2014). A hybrid algorithm for Bayesian network structure learning with application to multi-label learning.

[5] C. Bielza, G. Li, and P. Larrañaga (2011). Multi-dimensional classification with Bayesian networks.

[6] M. Gasse and A. Aussem (2016). Identifying the irreducible disjoint factors of a multivariate probability distribution.

Irreducible Label Factors

Theoretical results

Algebraic structure: if $\mathbf{Y}_{F_i}$ and $\mathbf{Y}_{F_j}$ are two LFs, then

- $\mathbf{Y}_{F_i} \cup \mathbf{Y}_{F_j}$ is a LF;
- $\mathbf{Y}_{F_i} \cap \mathbf{Y}_{F_j}$ is a LF;
- $\mathbf{Y}_{F_i} \setminus \mathbf{Y}_{F_j}$ is a LF.

Theoretical results

Algebraic structure: if $\mathbf{Y}_{F_i}$ and $\mathbf{Y}_{F_j}$ are two LFs, then

- $\mathbf{Y}_{F_i} \cup \mathbf{Y}_{F_j}$ is a LF;
- $\mathbf{Y}_{F_i} \cap \mathbf{Y}_{F_j}$ is a LF;
- $\mathbf{Y}_{F_i} \setminus \mathbf{Y}_{F_j}$ is a LF.

$\Longrightarrow$ the decomposition of $\mathbf{Y}$ into ILFs is unique.

Theoretical results

Algebraic structure: if $\mathbf{Y}_{F_i}$ and $\mathbf{Y}_{F_j}$ are two LFs, then

- $\mathbf{Y}_{F_i} \cup \mathbf{Y}_{F_j}$ is a LF;
- $\mathbf{Y}_{F_i} \cap \mathbf{Y}_{F_j}$ is a LF;
- $\mathbf{Y}_{F_i} \setminus \mathbf{Y}_{F_j}$ is a LF.

$\implies$ the decomposition of $\mathbf{Y}$ into ILFs is unique.

Constraint-based characterization:

- identifying all ILFs requires $O(m^2)$ pairwise CI tests;
- a practical procedure under the Composition assumption.

Theoretical results

Quadratic testing

Theorem

$<$ *any strict total order of* $\mathbf{Y}$.

$\mathcal{G} \leftarrow (\mathbf{Y}, \emptyset)$ *(empty graph)*
for all $Y_i \in \mathbf{Y}$ ***do***
    $\mathbf{Y}^i_{ind} \leftarrow \emptyset$
    for all $Y_j \in \{Y | Y > Y_i\}$ *(processed in* $<$ *order)* ***do***
        if $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \cup \{Y | Y < Y_i\} \cup \mathbf{Y}^i_{ind}$ ***then***
            $\mathbf{Y}^i_{ind} \leftarrow \mathbf{Y}^i_{ind} \cup \{Y_j\}$
        else
            Insert a new edge (i, j) *in* $\mathcal{G}$

$\Longrightarrow$ *each connected component is an ILF.*

Theoretical results

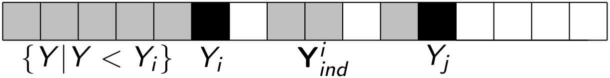
Quadratic testing

Theorem

$<$ any strict total order of $\mathbf{Y}$.

$\mathcal{G} \leftarrow (\mathbf{Y}, \emptyset)$ *(empty graph)*
for all $Y_i \in \mathbf{Y}$ **do**
    $\mathbf{Y}^i_{ind} \leftarrow \emptyset$
    for all $Y_j \in \{Y | Y > Y_i\}$ *(processed in $<$ order)* **do**
        if $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \cup \{Y | Y < Y_i\} \cup \mathbf{Y}^i_{ind}$ **then**
            $\mathbf{Y}^i_{ind} \leftarrow \mathbf{Y}^i_{ind} \cup \{Y_j\}$
        else
            Insert a new edge (i, j) in $\mathcal{G}$

$\implies$ *each connected component is an ILF.*



Theoretical results

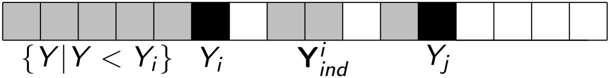
Quadratic testing

Theorem

$<$ any strict total order of $\mathbf{Y}$.

$\mathcal{G} \leftarrow (\mathbf{Y}, \emptyset)$ *(empty graph)*
for all $Y_i \in \mathbf{Y}$ ***do***
    $\mathbf{Y}^i_{ind} \leftarrow \emptyset$
    for all $Y_j \in \{Y | Y > Y_i\}$ *(processed in $<$ order)* ***do***
        if $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \cup \{Y | Y < Y_i\} \cup \mathbf{Y}^i_{ind}$ ***then***
            $\mathbf{Y}^i_{ind} \leftarrow \mathbf{Y}^i_{ind} \cup \{Y_j\}$
        else
            Insert a new edge (i, j) in $\mathcal{G}$

$\Longrightarrow$ *each connected component is an ILF.*



Pros: no assumptions, $O(m^2)$ tests.
Cons: cascading effect, high dimensional tests.

Theoretical results

Assuming Composition

Theorem

$\mathcal{G} = (\mathbf{Y}, \mathcal{E})$ an undirected graph, $Y_i - Y_j$ iff $Y_i \not\perp\!\!\!\perp Y_j \mid \mathbf{X}$
$\underset{compo}{\Longrightarrow}$ each connected component is an ILF.

Theoretical results

Assuming Composition

Theorem

$\mathcal{G} = (\mathbf{Y}, \mathcal{E})$ an undirected graph, $Y_i - Y_j$ iff $Y_i \not\perp\!\!\!\perp Y_j \mid \mathbf{X}$
$\underset{compo}{\Longrightarrow}$ each connected component is an ILF.

Moreover: $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \underset{compo}{\Longleftrightarrow} Y_i \perp\!\!\!\perp Y_j \mid \mathbf{M}_i$
with $\mathbf{M}_i$ a Markov boundary (minimum feature subset) of Y_i in $\mathbf{X}$.

Theoretical results

Assuming Composition

Theorem

$\mathcal{G} = (\mathbf{Y}, \mathcal{E})$ an undirected graph, $Y_i - Y_j$ iff $Y_i \not\perp\!\!\!\perp Y_j \mid \mathbf{X}$
$\underset{compo}{\Longrightarrow}$ each connected component is an ILF.

Moreover: $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \underset{compo}{\Longleftrightarrow} Y_i \perp\!\!\!\perp Y_j \mid \mathbf{M}_i$
with $\mathbf{M}_i$ a Markov boundary (minimum feature subset) of Y_i in $\mathbf{X}$.

Even better: $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \underset{compo}{\Longleftrightarrow} Y_i \perp\!\!\!\perp Y_j \mid S_i$
with $s_i = p(y_i|\mathbf{x})$ (a.k.a. propensity score).

Theoretical results

Assuming Composition

Theorem

$\mathcal{G} = (\mathbf{Y}, \mathcal{E})$ an undirected graph, $Y_i - Y_j$ iff $Y_i \not\perp\!\!\!\perp Y_j \mid \mathbf{X}$
$\underset{compo}{\Longrightarrow}$ each connected component is an ILF.

Moreover: $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \underset{compo}{\Longleftrightarrow} Y_i \perp\!\!\!\perp Y_j \mid \mathbf{M}_i$
with $\mathbf{M}_i$ a Markov boundary (minimum feature subset) of Y_i in $\mathbf{X}$.

Even better: $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \underset{compo}{\Longleftrightarrow} Y_i \perp\!\!\!\perp Y_j \mid S_i$
with $s_i = p(y_i|\mathbf{x})$ (a.k.a. propensity score).

Pros: $O(m^2)$ tests, low-dimensional.
Cons: Composition assumption.

Theoretical results

Assuming Composition

Dependency of a whole implies dependency of some part,

$$A \not\perp\!\!\!\perp \{B, C\} \mid D \implies A \not\perp\!\!\!\perp B \mid D \text{ or } A \not\perp\!\!\!\perp C \mid D.$$

Theoretical results

Assuming Composition

Dependency of a whole implies dependency of some part,

$$A \not\perp\!\!\!\perp \{B, C\} \mid D \implies A \not\perp\!\!\!\perp B \mid D \text{ or } A \not\perp\!\!\!\perp C \mid D.$$

Counter-example: $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$, $\mathbf{X} = \emptyset$, XOR relationship
$p(Y_1 = Y_2 \oplus Y_3) = \alpha$
$\{Y_1\} \not\perp\!\!\!\perp \{Y_2, Y_3\}$, yet $Y_1 \perp\!\!\!\perp Y_2$ and $Y_1 \perp\!\!\!\perp Y_3$.

Theoretical results

Assuming Composition

Dependency of a whole implies dependency of some part,

$$A \not\perp\!\!\!\perp \{B, C\} \mid D \implies A \not\perp\!\!\!\perp B \mid D \text{ or } A \not\perp\!\!\!\perp C \mid D.$$

Counter-example: $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$, $\mathbf{X} = \emptyset$, XOR relationship
$p(Y_1 = Y_2 \oplus Y_3) = \alpha$
$\{Y_1\} \not\perp\!\!\!\perp \{Y_2, Y_3\}$, yet $Y_1 \perp\!\!\!\perp Y_2$ and $Y_1 \perp\!\!\!\perp Y_3$.

Weak assumption, many approaches assume Composition:

- Linear models, multivariate Gaussian models;
- Greedy PGM structure learning algorithms (edge addition);
- Greedy FSS procedures (forward selection).

Theoretical results

Assuming Composition

Dependency of a whole implies dependency of some part,

$$A \not\perp\!\!\!\perp \{B, C\} \mid D \implies A \not\perp\!\!\!\perp B \mid D \text{ or } A \not\perp\!\!\!\perp C \mid D.$$

Counter-example: $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$, $\mathbf{X} = \emptyset$, XOR relationship
$p(Y_1 = Y_2 \oplus Y_3) = \alpha$
$\{Y_1\} \not\perp\!\!\!\perp \{Y_2, Y_3\}$, yet $Y_1 \perp\!\!\!\perp Y_2$ and $Y_1 \perp\!\!\!\perp Y_3$.

Weak assumption, many approaches assume Composition:

- Linear models, multivariate Gaussian models;
- Greedy PGM structure learning algorithms (edge addition);
- Greedy FSS procedures (forward selection).

My favorite: XOR is the basis of cryptography.

Theoretical results

Assuming Composition

Efficient procedure: ILF-Compo

1. for each label Y_i
   - learn $p(y_i \mid \mathbf{x})$ (probabilistic model);
   - obtain the propensity score s_i of each observation;
   - make s_i discrete (quantile discretization);
2. for each pair (Y_i, Y_j)
   - measure $Y_i \perp\!\!\!\perp Y_j \mid S_i$ and $Y_i \perp\!\!\!\perp Y_j \mid S_j$ (statistical tests);
   - place $Y_i - Y_j$ in $\mathcal{G}$ accordingly;
3. read connected components in $\mathcal{G}$ (breadth-first-search).

Experiments

MLC decomposition under $L_{0/1}$:

$$\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg_{\mathbf{y}} \prod_{k=1}^{n} \max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x}).$$

Experiments

MLC decomposition under $L_{0/1}$:

$$\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg_{\mathbf{y}} \prod_{k=1}^{n} \max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x}).$$

We compare three classification schemes

- **LP** (Label Powerset): $\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$
  - 1 classifier, 2^m classes (much less in practice)

Experiments

MLC decomposition under $L_{0/1}$:

$$\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg_{\mathbf{y}} \prod_{k=1}^{n} \max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x}).$$

We compare three classification schemes

- **LP** (Label Powerset): $\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$
  - 1 classifier, 2^m classes (much less in practice)
- **F-LP** (ILF-Compo + LP): $\arg\max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x})$ for each ILF
  - n classifiers, 2^{m_k} classes each

Experiments

MLC decomposition under $L_{0/1}$:

$$\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg_{\mathbf{y}} \prod_{k=1}^{n} \max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x}).$$

We compare three classification schemes

- **LP** (Label Powerset): $\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$
  - 1 classifier, 2^m classes (much less in practice)
- **F-LP** (ILF-Compo + LP): $\arg\max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x})$ for each ILF
  - n classifiers, 2^{m_k} classes each
- **BR** (Binary Relevance): $\arg\max_{y_i} p(y_i|\mathbf{x})$ for each label
  - m binary classifiers

Experiments

MLC decomposition under $L_{0/1}$:

$$\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) = \arg_{\mathbf{y}} \prod_{k=1}^{n} \max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x}).$$

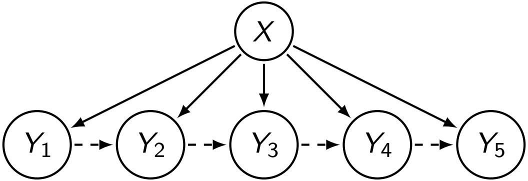
We compare three classification schemes

- **LP** (Label Powerset): $\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$
  - 1 classifier, 2^m classes (much less in practice)
- **F-LP** (ILF-Compo + LP): $\arg\max_{\mathbf{y}_{F_k}} p(\mathbf{y}_{F_k}|\mathbf{x})$ for each ILF
  - n classifiers, 2^{m_k} classes each
- **BR** (Binary Relevance): $\arg\max_{y_i} p(y_i|\mathbf{x})$ for each label
  - m binary classifiers

Same base learner.

Experiments

Synthetic toy problem



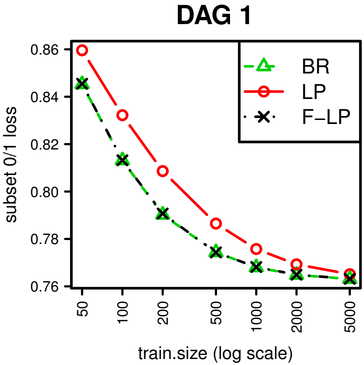
Generic toy DAG (Bayesian network).

We build 5 distinct factorizations:

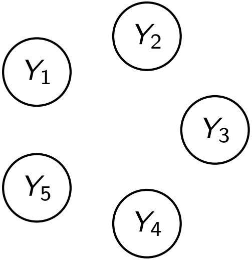
- DAG 1: $\{Y_1\}, \{Y_2\}, \{Y_3\}, \{Y_4\}, \{Y_5\}$;
- DAG 2: $\{Y_1, Y_2\}, \{Y_3, Y_4\}, \{Y_5\}$;
- DAG 3: $\{Y_1, Y_2, Y_3\}, \{Y_4, Y_5\}$;
- DAG 4: $\{Y_1, Y_2, Y_3, Y_4\}, \{Y_5\}$;
- DAG 5: $\{Y_1, Y_2, Y_3, Y_4, Y_5\}$.

Experiments

Synthetic toy problem



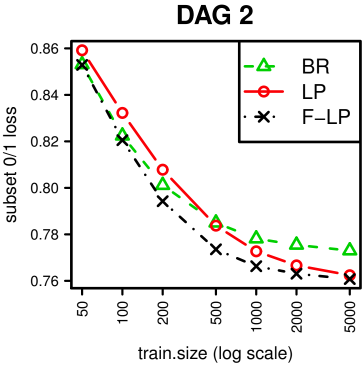
Test set $L_{0/1}$ over 1000 runs.



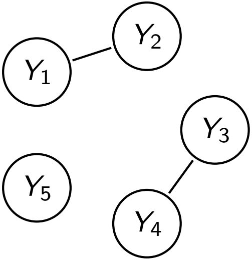
Decomposition graph.

Experiments

Synthetic toy problem



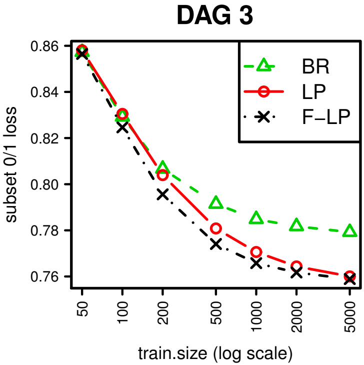
Test set $L_{0/1}$ over 1000 runs.



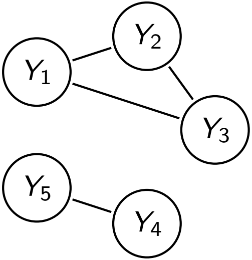
Decomposition graph.

Experiments

Synthetic toy problem



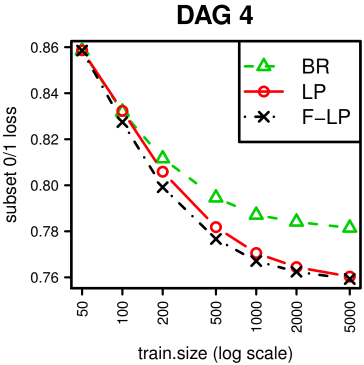
Test set $L_{0/1}$ over 1000 runs.



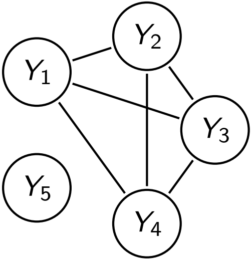
Decomposition graph.

Experiments

Synthetic toy problem



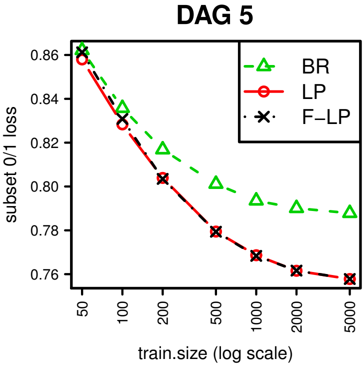
Test set $L_{0/1}$ over 1000 runs.



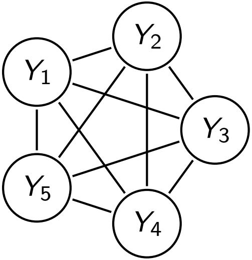
Decomposition graph.

Experiments

Synthetic toy problem



Test set $L_{0/1}$ over 1000 runs.



Decomposition graph.

Experiments

Real-world data sets

8 standard multi-label data sets[7]

dataset	domain	$\lvert\mathcal{D}\rvert$	$\lvert\mathbf{X}\rvert$	$\lvert\mathbf{Y}\rvert$
emotions	music	593	72	6
image	images	2000	135	5
scene	images	2407	294	6
yeast	biology	2417	103	14
slashdot	text	3782	1079	22
genbase	biology	662	1186	27
medical	text	978	1449	45
enron	text	1702	1001	53

[7] http://mulan.sourceforge.net/datasets-mlc.html

Experiments

Real-world data sets

8 standard multi-label data sets[7]

dataset	domain	$\lvert\mathcal{D}\rvert$	$\lvert\mathbf{X}\rvert$	$\lvert\mathbf{Y}\rvert$
emotions	music	593	72	6
image	images	2000	135	5
scene	images	2407	294	6
yeast	biology	2417	103	14
slashdot	text	3782	1079	22
genbase	biology	662	1186	27
medical	text	978	1449	45
enron	text	1702	1001	53

7 additional MLC approaches:

- CC, PCC, MCC, ECC, RAkEL, HOMER, LEAD

[7] http://mulan.sourceforge.net/datasets-mlc.html

Experiments

Real-world data sets

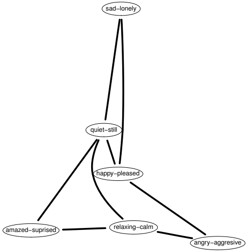
Mean $L_{0/1}$ over 5x2 cv (lower is better):

method	emotions	image	scene	yeast	slashdot	genbase	medical	enron
LP	**66.2**	**53.7**	**31.5**	**75.1**	**55.0**	3.8	33.0	**83.8**
F-LP	**66.2**	**53.7**	31.8	**75.1**	59.1	3.4	32.2	85.3
BR	73.6	76.4	49.0	85.5	66.2	3.4	35.9	89.3
CC	71.6	57.9	37.0	80.7	62.0	3.3	32.7	88.0
ECC	70.6	59.7	37.7	79.8	60.3	**3.1**	**31.7**	86.9
MCC	67.9	57.3	37.2	79.8	61.9	3.4	33.4	88.1
PCC	70.7	59.7	39.8	79.6	-	-	-	-
RAkEL	69.3	57.8	39.4	81.6	65.3	3.2	35.6	89.0
HOMER	71.7	68.4	49.4	86.9	64.9	3.4	37.9	89.7
LEAD	76.2	70.2	49.9	85.4	69.2	3.8	37.4	91.8

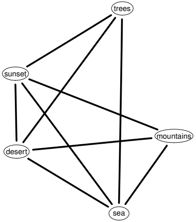
win / tie / loss = 2 / 3 / 3

Experiments

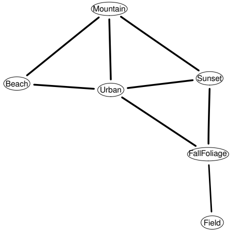
Real-world data sets



emotions

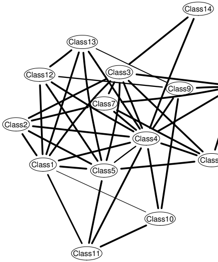


image



scene

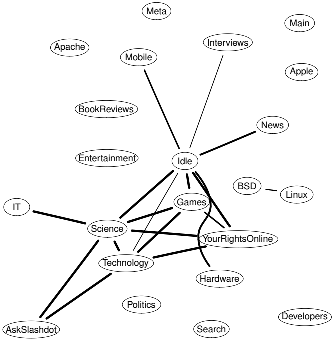
F-LP = LP



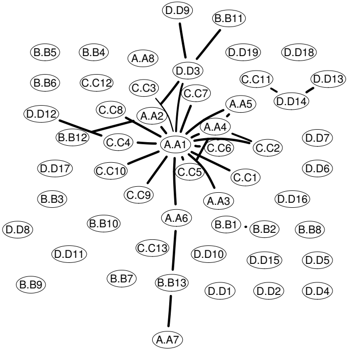
yeast

Experiments

Real-world data sets



slashdot

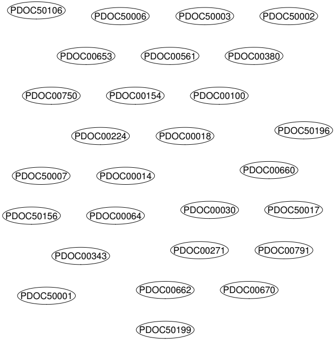


enron

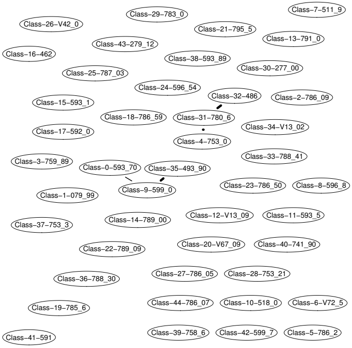
F-LP $<$ LP

Experiments

Real-world data sets



genbase

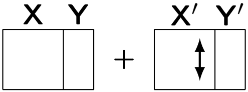


medical

F-LP > LP

Experiments

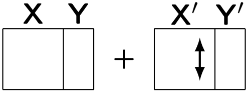
Twin data sets



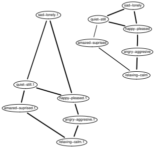
$$\Longrightarrow \mathbf{Y} \perp\!\!\!\perp \mathbf{Y}' \mid \mathbf{X} \cup \mathbf{X}'$$

Experiments

Twin data sets



$$\Longrightarrow \mathbf{Y} \perp\!\!\!\perp \mathbf{Y}' \mid \mathbf{X} \cup \mathbf{X}'$$



emotions2

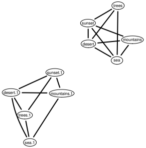
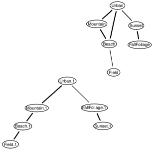
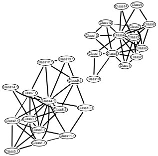


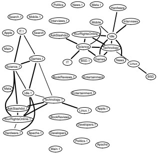
image2



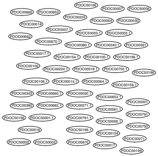
scene2



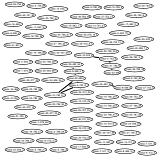
yeast2



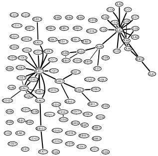
slashdot2



genbase2



medical2



enron2

Mean $L_{0/1}$ over 5x2 cv (lower is better):

method	emotions2	image2	scene2	yeast2	slashdot2	genbase2	medical2	enron2
LP	94.9	87.6	62.8	97.5	90.3	33.7	86.6	99.3
F-LP	**91.8**	**82.0**	**58.6**	**95.0**	**83.9**	**6.8**	**62.4**	**98.4**
BR	94.7	93.7	79.0	98.0	89.9	**6.8**	67.0	99.1
CC	95.1	83.9	66.9	96.5	86.5	7.1	64.4	99.0
ECC	93.6	84.8	66.5	97.0	86.1	7.2	64.4	98.7
MCC	93.6	85.6	67.9	96.4	86.6	7.1	64.4	98.9
PCC	93.1	85.9	71.0	-	-	-	-	-
RAkEL	93.7	89.7	72.0	97.8	89.3	**6.8**	67.2	99.2
HOMER	95.5	91.8	79.9	98.8	97.0	27.0	82.1	99.6
LEAD	95.9	93.0	80.5	98.1	91.3	8.9	65.5	99.6

win / tie / loss = 10 / 0 / 0

Conclusions

Thesis in-between PGMs and MLC.

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for $0/1$ loss minimization (ICML 2015)

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for $0/1$ loss minimization (ICML 2015)
- F-GFM useful for F-measure maximization (ECML 2016)

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for $0/1$ loss minimization (ICML 2015)
- F-GFM useful for F-measure maximization (ECML 2016)

Limitations:

- a disjoint factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for 0/1 loss minimization (ICML 2015)
- F-GFM useful for F-measure maximization (ECML 2016)

Limitations:

- a disjoint factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- multiple testing (imagine two ILFs of size 100)

---

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for 0/1 loss minimization (ICML 2015)
- F-GFM useful for F-measure maximization (ECML 2016)

Limitations:

- a disjoint factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- multiple testing (imagine two ILFs of size 100)
- experimental results could be further improved

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for 0/1 loss minimization (ICML 2015)
- F-GFM useful for F-measure maximization (ECML 2016)

Limitations:

- a disjoint factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- multiple testing (imagine two ILFs of size 100)
- experimental results could be further improved

Other contributions:

- H2PC for BN structure learning (ECML 2012, ESWA 2014)

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for 0/1 loss minimization (ICML 2015)
- F-GFM useful for F-measure maximization (ECML 2016)

Limitations:

- a disjoint factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- multiple testing (imagine two ILFs of size 100)
- experimental results could be further improved

Other contributions:

- H2PC for BN structure learning (ECML 2012, ESWA 2014)
- some conjectures on Chain Graphs

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Conclusions

Thesis in-between PGMs and MLC.

Main contribution: ILFs

- factorizing $p(\mathbf{y}|\mathbf{x})$ requires $O(m^2)$ CI tests (PGM 2016)
- F-LP useful for 0/1 loss minimization (ICML 2015)
- F-GFM useful for F-measure maximization (ECML 2016)

Limitations:

- a disjoint factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- multiple testing (imagine two ILFs of size 100)
- experimental results could be further improved

Other contributions:

- H2PC for BN structure learning (ECML 2012, ESWA 2014)
- some conjectures on Chain Graphs
- SPNlearn[8] factorization optimal under Composition

[8] H. Poon and P. M. Domingos (2011). Sum-Product Networks: A New Deep Architecture.

Perspectives

Score-based approach

- score-based structures usually more consistent
- $O(m^2)$ CI characterization $\implies$ $O(m^2)$ search strategy?

[9] M. Studeny (2005). Probabilistic Conditional Independence Structures.

Perspectives

Score-based approach

- score-based structures usually more consistent
- $O(m^2)$ CI characterization $\implies$ $O(m^2)$ search strategy?

Representation learning

- factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- learn $\mathbf{z} = \mathbf{f}(\mathbf{y})$ such that $p(\mathbf{z}|\mathbf{x})$ factorizes

[9] M. Studeny (2005). Probabilistic Conditional Independence Structures.

Perspectives

Score-based approach

- score-based structures usually more consistent
- $O(m^2)$ CI characterization $\implies$ $O(m^2)$ search strategy?

Representation learning

- factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- learn $\mathbf{z} = \mathbf{f}(\mathbf{y})$ such that $p(\mathbf{z}|\mathbf{x})$ factorizes

Decomposable models

- CI characterization still open problem[9]
- non-disjoint factorization generalizes ILFs

[9] M. Studeny (2005). Probabilistic Conditional Independence Structures.

Perspectives

Score-based approach

- score-based structures usually more consistent
- $O(m^2)$ CI characterization $\implies$ $O(m^2)$ search strategy?

Representation learning

- factorization of $p(\mathbf{y}|\mathbf{x})$ is not guaranteed
- learn $\mathbf{z} = \mathbf{f}(\mathbf{y})$ such that $p(\mathbf{z}|\mathbf{x})$ factorizes

Decomposable models

- CI characterization still open problem[9]
- non-disjoint factorization generalizes ILFs

Post-doc: deep learning for image inpainting (CREATIS)

[9] M. Studeny (2005). Probabilistic Conditional Independence Structures.

Probabilistic Graphical Model Structure Learning: Application to Multi-Label Classification

PhD defense

Maxime Gasse

Supervised by: Alex Aussem and Haytham Elghazel

Thank you!

Proof: propensity score

$s_i = p(y_i \mid \mathbf{x})$
captures all - and only - information from $\mathbf{X}$ about Y_i:
$Y_i \perp\!\!\!\perp \mathbf{X} \mid S_i$ and $Y_i \perp\!\!\!\perp S_i \mid \mathbf{X}$.

$$
\begin{aligned}
& Y_i \perp\!\!\!\perp Y_j \mid S_i & \\
\Longrightarrow\ & Y_i \perp\!\!\!\perp Y_j \cup \mathbf{X} \mid S_i & \text{(Composition with } Y_i \perp\!\!\!\perp \mathbf{X} \mid S_i\text{)} \\
\Longrightarrow\ & Y_i \perp\!\!\!\perp Y_j \mid S_i \cup \mathbf{X} & \text{(Weak Union)} \\
\Longrightarrow\ & Y_i \perp\!\!\!\perp Y_j, S_i \mid \mathbf{X} & \text{(Contraction with } Y_i \perp\!\!\!\perp S_i \mid \mathbf{X}\text{)} \\
\Longrightarrow\ & Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} & \text{(Decomposition)}
\end{aligned}
$$

$$Y_i \perp\!\!\!\perp Y_j \mid S_i \underset{compo}{\Longrightarrow} Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X}$$

The demonstration $Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X} \underset{compo}{\Longrightarrow} Y_i \perp\!\!\!\perp Y_j \mid S_i$ is the same.

Experiments

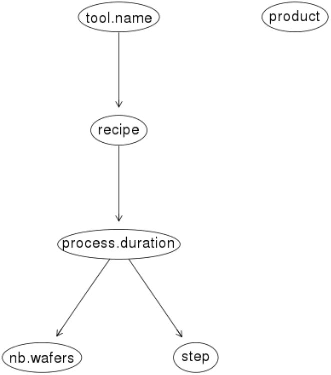
Varying α

Mean $L_{0/1}$ over 5x2 cv (lower is better):

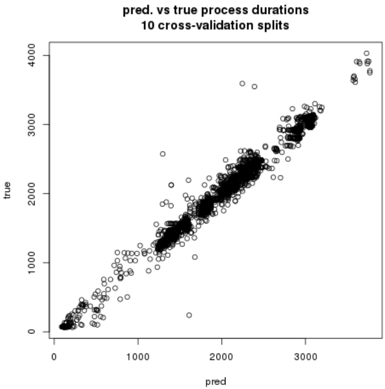
method	emotions	image	scene	yeast	slashdot	genbase	medical	enron
LP	**67.6**	**53.5**	**31.8**	75.2	**56.0**	3.5	**32.4**	**83.9**
F-LP ($\alpha = 10^{-1}$)	**67.6**	**53.5**	**31.8**	75.2	**56.0**	3.6	**32.4**	**83.9**
F-LP ($\alpha = 10^{-2}$)	**67.6**	**53.5**	**31.8**	75.2	**56.0**	3.4	32.8	**83.9**
F-LP ($\alpha = 10^{-4}$)	**67.6**	**53.5**	**31.8**	75.2	56.5	3.7	33.5	85.2
F-LP ($\alpha = 10^{-8}$)	68.4	**53.5**	**31.8**	75.2	61.7	3.2	35.1	86.8
F-LP ($\alpha = 10^{-16}$)	73.7	57.3	32.6	**75.1**	66.0	**2.9**	35.8	88.3
BR	73.9	76.0	48.7	85.8	66.6	**2.9**	35.8	89.2

STMicroelectronics

Use case: process duration



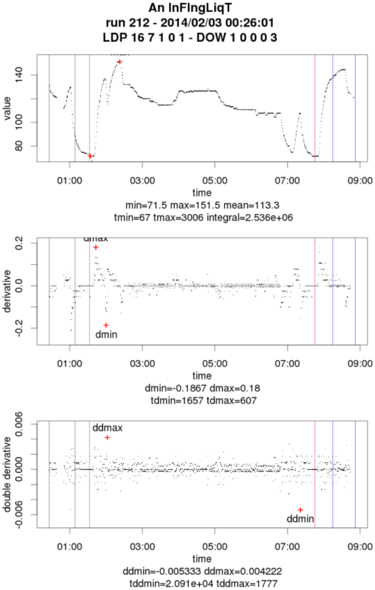
structure learning

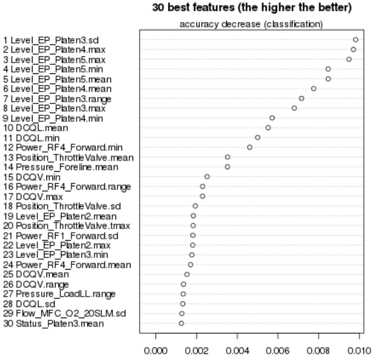


regression

STMicroelectronics

Use case: wafer contamination





feature extraction