Introduction

Probabilistic Graphical Models

Jerónimo Hernández-González

Revisiting concepts

- What is a Probabilistic Graphical Model?
- Which types of PGMs do you know?
- Which are the differences between them?
- Why do we use PGMs?

The Artificial Intelligence problem simplified

Given a problem,	Flu diagnosis
we use a model to represent it	train a classifier
so that we can ask questions to the model	predict (help to diagnose) a new patient with flu

The logical approach to AI

Given a problem,

1. Representation:
   a) Determine which are the facts and rules that are relevant to the problem.
   b) Represent them by means of logical theory.
2. Inference:
   a) Rewrite your question as a logic formula such that the answer (true/false) to that formula is your answer.
   b) Determine whether the formula is satisfied or not using logical inference (modus ponens, modus tollens, ...).

The logical approach to AI

Given a problem,

1. Representation:
   a) Determine which are the facts and rules that are relevant to the problem.
   b) Represent them by means of logical theory.
2. Inference:
   a) Rewrite your question as a logic formula such that the answer (true/false) to that formula is your answer.
   b) Determine whether the formula is satisfied or not using logical inference (modus ponens, modus tollens, ...).

Most of the time, things are not just true or false...

The probabilistic approach to AI

Given a problem which involves uncertainty,

1. Representation.
   a) Identify the relevant variables (observable or not) related to your question
   b) Define a probability distribution over all the variables identified.
2. Inference.
   a) Rewrite your question as a probabilistic query such that your answer is a distribution over possible outcomes.
   b) Obtain the probability distribution by using probabilistic inference.

Example I: Medical diagnosis

In a healthcare center, a patient comes in coughing.
A physician wants to determine how likely it is that she has flu.

1. Representation.

   a) We identify the relevant variables:
      1. whether the patient coughs $C = \{T, F\}$
      2. whether the patient has fever $F = \{T, F\}$
      3. whether the patient has a flu $I = \{T, F\}$

   b) We define the distribution $P(C, F, I)$ with the help of the physician, who tells the probability of every combination of (C, F, I)

2. Inference.

   a) Our question, rewritten as a probabilistic query, is:
      Which is the probability distribution, $P(I|C = T)$?

   b) We obtain the probability distribution over I using inference.

Example II: Medical diagnosis

In a healthcare center, a second patient comes in. She is coughing and suffers hemoptysis.
A physician wants to know whether she has flu, lung cancer or none.

1. Representation. `Need to improve the model`
   a) We identify the relevant variables:
      1 C and F as in the previous example
      2 whether the patient shows hemoptysis $M = \{T, F\}$
      3 patient's diagnosis $D = \{Cancer, Flu, None\}$
   b) We define the dist. $P(C, F, M, D)$ with the help of the physician, who tells the probability of every combination of (C, F, M, D)
2. Inference. `A different type of query`
   a) Our question, rewritten as a probabilistic query, is:
      Which is the diagnosis $\arg \text{máx}_d P(D = d | C = T, M = T)$?
   b) We use inference to obtain d.

Types of probabilistic queries

Given a probability distribution $P(\boldsymbol{X}) = P(X_1, \ldots, X_n)$,

a partition of the set of variables $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$,

and an value-assignment $\boldsymbol{e}$ to the subset of variables $\boldsymbol{E}$,

we identify **two different types** of probabilistic queries:

1. Conditional probability queries. Find the *marginal* distribution:

$$P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$$

2. MAP queries. Find the value-assignment $\boldsymbol{y}$ to the subset $\boldsymbol{Y}$ that maximizes $P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{E} = \boldsymbol{e})$:

$$\arg \max_{\boldsymbol{y}} P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{E} = \boldsymbol{e})$$

The probabilistic approach to AI ***without experts!***

Given a problem which involves uncertainty and data,

1. Representation.
   a) Identify the relevant variables (observable or not) related to your question
   b) Define the set of possible probability distributions over all the variables identified.
2. Learning.
   Based on the available data, select the single probability distribution in the set that is more likely.
3. Inference.
   a) Rewrite your question as a probabilistic query such that your answer is a distribution over possible outcomes.
   b) Obtain the probability distribution by using probabilistic inference.

The probabilistic approach to AI ***without experts!***

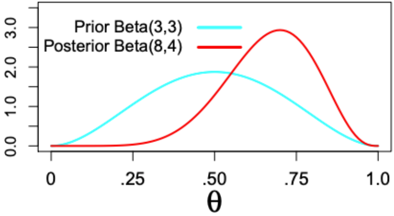
Given a problem which involves uncertainty, data and initial beliefs,

1. Representation.
   a) Identify the relevant variables (known and unknown) related to your question
   b) Define the set of possible prob. distributions over all the variables identified and weigh them w.r.t. your initial belief.
2. Learning.
   With the available data, refine your initial beliefs weighing more the probability distributions in the set that are more likely
3. Inference.
   a) Rewrite your question as a probabilistic query such that your answer is a distribution over possible outcomes.
   b) Obtain the probability distribution by using probabilistic inference (a weighted combination of all prob. distributions).

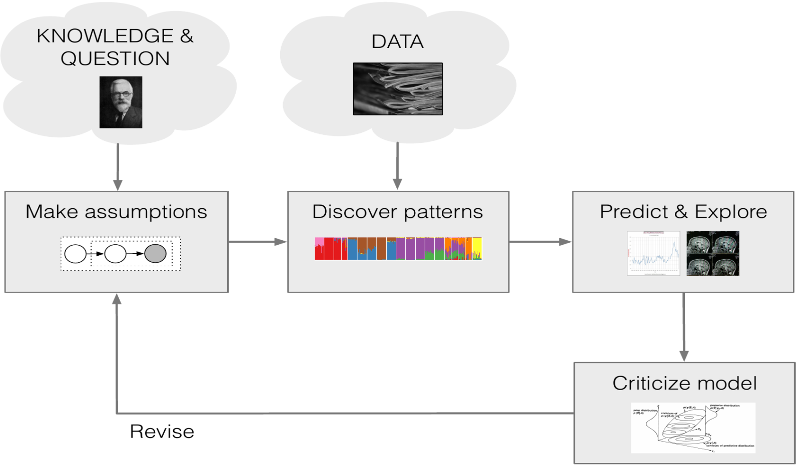
The probabilistic approach to AI ***without experts!***

Examples first case: A novel physician learns from the records of a retired doctor.
Frequentist coin example (HTTTTT): $\theta = 5/6$

Examples second case: A novel physician updates her academic beliefs with the records of a retired doctor.
Bayesian coin example (HTTTTT): $p(\theta)$



The probabilistic approach to data science



[Box, 1980; Rubin, 1984; Gelman+ 1996; Blei, 2014]

The probabilistic approach to AI: drawbacks

Example: a genetic engineer

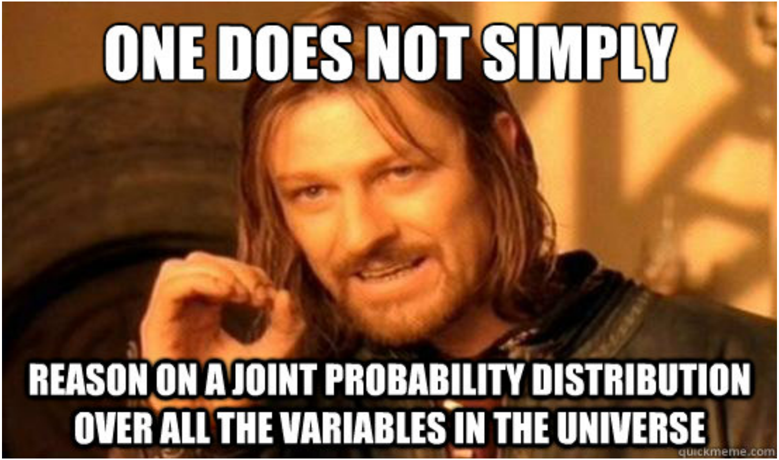
She wants to deal with nucleotide sequences.
We identify n relevant variables (one per nucleotide) with 4 possible values each $\boldsymbol{X} = (X_1, \ldots, X_n)$:

- How much memory does the joint distribution $P(X)$ need?

$$4^n$$

- What is the cost of finding the most probable assignment of values to variables?

The probabilistic approach to AI: drawbacks



Probabilistic Graphical Models

We need a way to:

- Encode probability distributions over large number of variables in a compact way
- Efficiently answer queries
- Learn from the available data

Probabilistic Graphical Models

We need a way to:

- Encode probability distributions over large number of variables in a compact way
- Efficiently answer queries
- Learn from the available data

PGMs do the job!

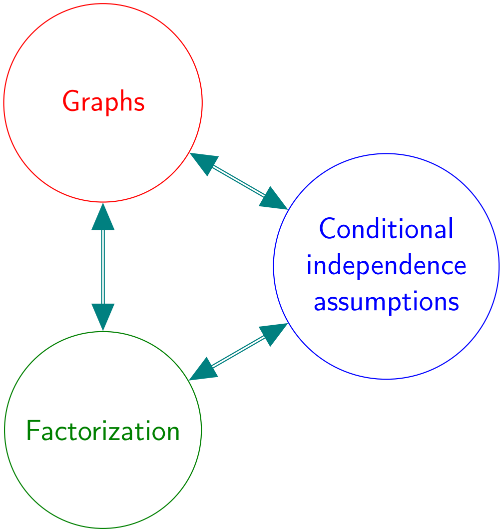
- Use a graph:

  Each variable is represented by a vertex in the graph. *Possible dependencies* between variables are encoded by adding edges to the graph
- Why graphs?
  - Easy to visualize and understand.
  - Mathematically adequate.

Probabilistic graphical models

- Given by:
  - Structure: a graph
  - Parameters: a set of conditional probabilities
- Three main types:
  - Directed acyclic graphs: Bayesian networks (a.k.a. belief networks)
  - Undirected graphs: Markov networks (a.k.a. Markov random fields)
  - Bipartite graphs: Factor graphs
- They represent:
  - Factorizations of probability distributions: Chain rule with a reduced set of conditioning variables
  - Dependency models: A set of conditional independences

Three views of probabilistic modeling



Probabilistic independence

2 random variables are independent if, for all values that Z and Y can take,

$$P(Y, Z) = P(Y) \cdot P(Z)$$

An equivalent condition is:

$$P(Y|Z) = P(Y) \quad \text{or} \quad P(Z|Y) = P(Z)$$

We note it as $\quad Y \perp\!\!\!\perp Z$

** Same definition when $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of variables.

Probabilistic independence

Examples

The probability that there is electricity in this room today is...

- **independent** of whether Manchester City wins European Champions League this year.
- **not** independent of whether there was electricity yesterday.
- **not** independent of whether there is electricity in the room next door.
- **independent** of whether I took a shower this morning.

*Probabilistic **conditional** independence*

Given 3 random variables, X, Y, Z we say that X is conditionally independent of Y given Z if,

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

We note it as $\quad X \perp\!\!\!\perp Y|Z.$

** Same definition when $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of variables.

When modeling a problem, some conditional independence relations are usually clear.

Probabilistic conditional independence

Examples

- Kevin separately phones two students, Alice and Bob, and tells both the same number, $n_k \in \{1, \ldots, 10\}$.
- Alice and Bob do not hear it well and each independently draw a conclusion about what Kevin said.
  Let n_a be the number Alice heard, and n_b the one Bob heard.
- Are n_a and n_b (marginally) independent?

Probabilistic conditional independence

Examples

- Kevin separately phones two students, Alice and Bob, and tells both the same number, $n_k \in \{1, \ldots, 10\}$.
- Alice and Bob do not hear it well and each independently draw a conclusion about what Kevin said.
  Let n_a be the number Alice heard, and n_b the one Bob heard.
- Are n_a and n_b (marginally) independent?
  No! We'd expect $P(n_a = 1 | n_b = 1) \neq P(n_a = 1)$

** n_b gives some evidence of what n_k might be

Probabilistic conditional independence

Examples

- Kevin separately phones two students, Alice and Bob, and tells both the same number, $n_k \in \{1, \ldots, 10\}$.
- Alice and Bob do not hear it well and each independently draw a conclusion about what Kevin said.
  Let n_a be the number Alice heard, and n_b the one Bob heard.
- Are n_a and n_b (marginally) independent?
  No! We'd expect $P(n_a = 1 | n_b = 1) \neq P(n_a = 1)$

  ** n_b gives some evidence of what n_k might be
- Are n_a and n_b conditionally independent given n_k?

Probabilistic conditional independence

Examples

- Kevin separately phones two students, Alice and Bob, and tells both the same number, $n_k \in \{1, \ldots, 10\}$.
- Alice and Bob do not hear it well and each independently draw a conclusion about what Kevin said.
  Let n_a be the number Alice heard, and n_b the one Bob heard.
- Are n_a and n_b (marginally) independent?
  No! We'd expect $P(n_a = 1 | n_b = 1) \neq P(n_a = 1)$
  ** n_b gives some evidence of what n_k might be
- Are n_a and n_b conditionally independent given n_k?
  Yes! If we know what Kevin actually said, n_a and n_b are no longer related.

$$P(n_a = 1 | n_b = 1, n_k = 2) = P(n_a = 1 | n_k = 2)$$

** n_k fully explains any possible relationships between n_a and n_b

Probabilistic conditional independence

Examples

- Two random variables that are marginally independent **can** also be conditionally independent given a third variable:

  The probability that [Manchester City wins this year Champions League] is independent of [whether there is electricity in this room] given that [there is electricity in the room next door].

- However that's not always the case:
  Given two coins C_1 and C_2, these are marginally independent

  $$P(C_1|C_2 = H) = P(C_1)$$

  Consider now a third variable: S is *true* if both coins show the same face

  $$P(C_1|C_2 = H, S = true) \neq P(C_1|S = true)$$

Independence: a modeling advantage

Independence is good for simplicity!

If we have to represent a probability distribution $P(\boldsymbol{X})$, and we know that we can split $\boldsymbol{X} = \boldsymbol{Y} \cup \boldsymbol{Z}$ such that

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z},$$

then we can represent $P(\boldsymbol{X}) = P(\boldsymbol{Y}) \cdot P(\boldsymbol{Z})$, that is, we need to represent just two smaller pieces, $P(\boldsymbol{Y})$ and $P(\boldsymbol{Z})$.

Example: Say $\boldsymbol{X}$ has 10 binary variables and $\boldsymbol{Y}$ and $\boldsymbol{Z}$ have 5 binary variables each. How much memory are we saving?

Independence: a modeling advantage

Independence is good for simplicity!

If we have to represent a probability distribution $P(\boldsymbol{X})$, and we know that we can split $\boldsymbol{X} = \boldsymbol{Y} \cup \boldsymbol{Z}$ such that

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z},$$

then we can represent $P(\boldsymbol{X}) = P(\boldsymbol{Y}) \cdot P(\boldsymbol{Z})$, that is, we need to represent just two smaller pieces, $P(\boldsymbol{Y})$ and $P(\boldsymbol{Z})$.

Example: Say $\boldsymbol{X}$ has 10 binary variables and $\boldsymbol{Y}$ and $\boldsymbol{Z}$ have 5 binary variables each. How much memory are we saving?

From 1024 to 64 parameters!

Structural statistical model

A set of (conditional) independence statements

Simple explanation with two variables, X and Y

There is a single independence statement that $P(X, Y)$ can satisfy:

$$X \perp\!\!\!\perp Y$$

There are two possible structural models:

1. $\mathcal{M}_1 = \emptyset$
2. $\mathcal{M}_2 = \{X \perp\!\!\!\perp Y\}$.

A distribution respects model $\mathcal{M}$ if it satisfies all the independencies in $\mathcal{M}$

All the possible prob. distributions $p(X, Y)$ can be classified into:

a) "Distributions that respect $\mathcal{M}_1$"
b) "Distributions that respect $\mathcal{M}_2$".

Structural statistical model

A set of (conditional) independence statements

Generalized explanation, with n variables

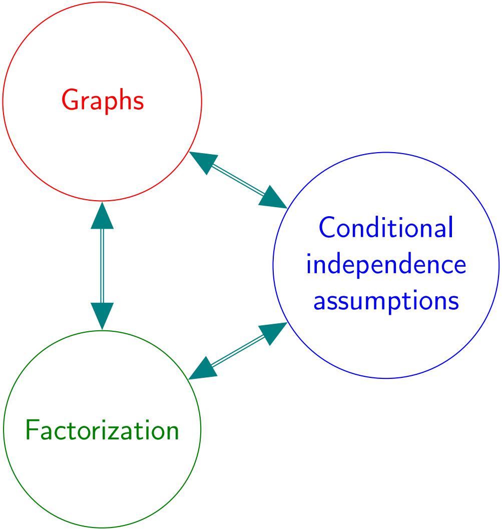
A structural statistical model over a set of variables $\boldsymbol{V}$ is described by a set of conditional independence assumptions like:

$$\begin{aligned} \mathcal{M} = \{ & \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}, \\ & \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} | \boldsymbol{Y}, \\ & \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \} \end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$.

A distribution respects model $\mathcal{M}$ if it satisfies all the independencies in $\mathcal{M}$

Three views of probabilistic modeling



Factorization

Example

Let us consider the following complex function:

$$f(x, y, z, t) = 6yzt + y^4 z - 6\sqrt{x}zt - \sqrt{x}y^3 z + 6yt \log t + y^4 t \log t - 6\sqrt{x}t \log t - \sqrt{x}y^3 \log t$$

Factorization

Example

Let us consider the following complex function:

$$f(x, y, z, t) = 6yzt + y^4z - 6\sqrt{x}zt - \sqrt{x}y^3z + 6yt\log t + y^4t\log t - 6\sqrt{x}t\log t - \sqrt{x}y^3\log t$$

It can be re-expressed as:

$$f(x, y, z, t) = (z + \log t) \cdot (y - \sqrt{x}) \cdot (6t + y^3)$$

Factorization

Example

Let us consider the following complex function:

$$f(x,y,z,t) = 6yzt + y^4 z - 6\sqrt{x}zt - \sqrt{x}y^3 z + 6yt\log t + y^4 t\log t - 6\sqrt{x}t\log t - \sqrt{x}y^3\log t$$

It can be re-expressed as:

$$f(x,y,z,t) = (z + \log t)\cdot(y - \sqrt{x})\cdot(6t + y^3)$$

We say that f factorizes as a product of three factors:

$$f(x,y,z,t) = f_1(z,t)\cdot f_2(x,y)\cdot f_3(y,t)$$

Factorization

Example

Note that the following factorization:

$$f(x, y, z, t) = f_1(z, t) \cdot f_2(x, y) \cdot f_3(y, t)$$

involves:

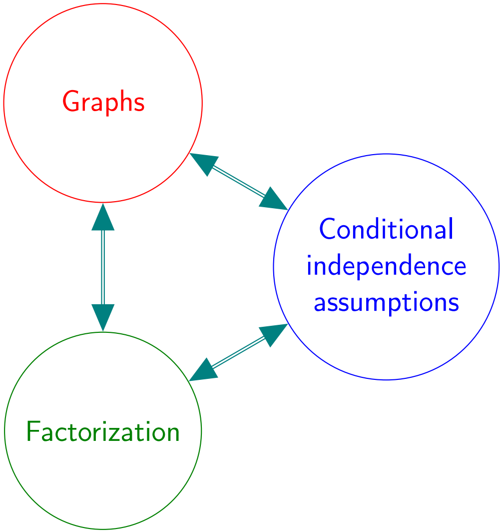
- $Scope(f) = \{x, y, z, t\}$
- $Scope(f_1) = \{z, t\}$
- $Scope(f_2) = \{x, y\}$
- $Scope(f_3) = \{z, t\}$

When a function (our prob. distribution) factorizes into small parts, we can represent it in a smaller space.

We are interested in probability distributions that factorize!!

Ex.: $P(X, Y, Z) = P(X|Z)P(Y|Z)P(Z)$

Three views of probabilistic modeling



Probabilistic graphical models

- Given by:
  - Structure: a graph
  - Parameters: a set of conditional probabilities
- Three main types:
  - Directed acyclic graphs: Bayesian networks (a.k.a. belief networks)
  - Undirected graphs: Markov networks (a.k.a. Markov random fields)
  - Bipartite graphs: Factor graphs
- They represent:
  - Factorizations of probability distributions: Chain rule with a reduced set of conditioning variables
  - Dependency models: A set of conditional independences

Distributions and factors

Approaches:

- Distributions
  - Joint distribution
  - Conditioning
  - Marginalization
- Factors
  - How are they different from distributions?
  - Product
  - Marginalization
  - Reduction

Inference

PGMs allow for probability based operations

- which can be done more efficiently (exact/approximate)
  - Exact: marginalize, conditioning, belief propagation,...
  - Approx.: random sampling
- in order to answer two types of queries

Learning

- From data (and expert knowledge)
- Learning algorithms for...
  - Parametric learning
  - Structural learning (NP-complete)
    - Quantitative: Scoring functions (e.g. likelihood)
    - Qualitative: Conditional independence tests
- To obtain robust models, we need to control the trade-off between model complexity and amount of available data

Relation with supervised learning

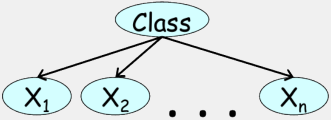
A generative approach

1. Learn the underlying statistical model
2. Classify unseen instances into the most probable class

** Bayes classifier: lower bound of the classification error

Structures biased towards supervised learning

Few parameters with high discriminative information
Ex.: (Augmented) naive Bayes family



What is all this about?

By the end of this course you should know:

- Representation: what is a probabilistic graphical model
- Inference: which queries can we ask to them
- Learning: how to obtain PGMs from data

What is all this about?

By the end of this course you should know:

- Representation: what is a probabilistic graphical model
  - Directed and Undirected
  - Temporal and plate models
- Inference: which queries can we ask to them
  - when (and how) these questions can be answered exactly in polynomial time (exact inference)
  - what to do when they cannot (approximate inference)
- Learning: how to obtain PGMs from data
  - Parameters and structure
  - With and without complete data

What is all this about?

By the end of this course you should know:

- Representation: what is a probabilistic graphical model
  - Directed and Undirected
  - Temporal and plate models
- Inference: which queries can we ask to them
  - when (and how) these questions can be answered exactly in polynomial time (exact inference)
  - what to do when they cannot (approximate inference)
- Learning: how to obtain PGMs from data
  - Parameters and structure
  - With and without complete data

Furthermore, you should be able to:

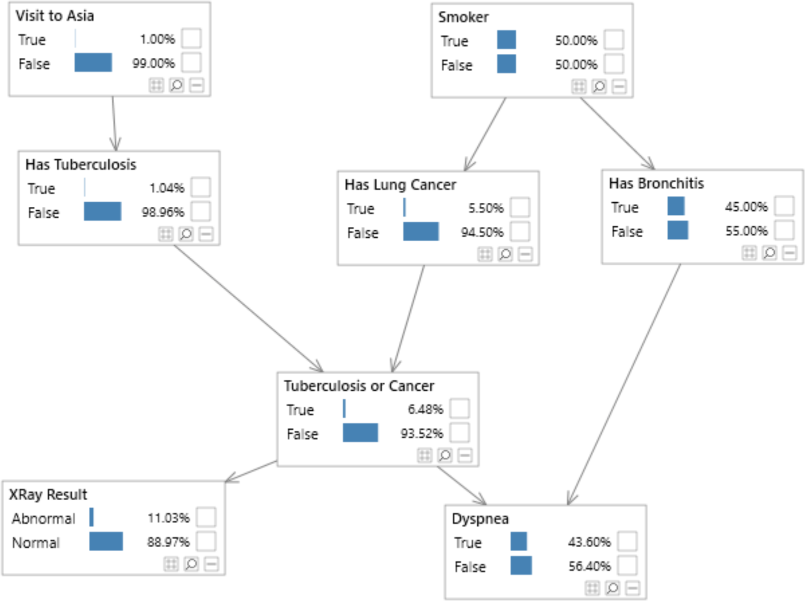
- apply PGM algorithms to problems of your interest
- translate PGMs and algorithms into code

Introduction

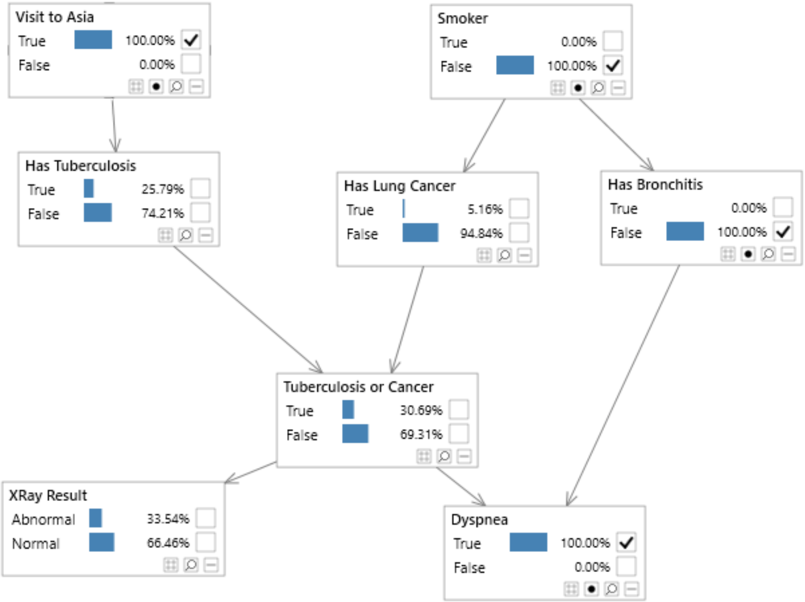
Probabilistic Graphical Models

Jerónimo Hernández-González

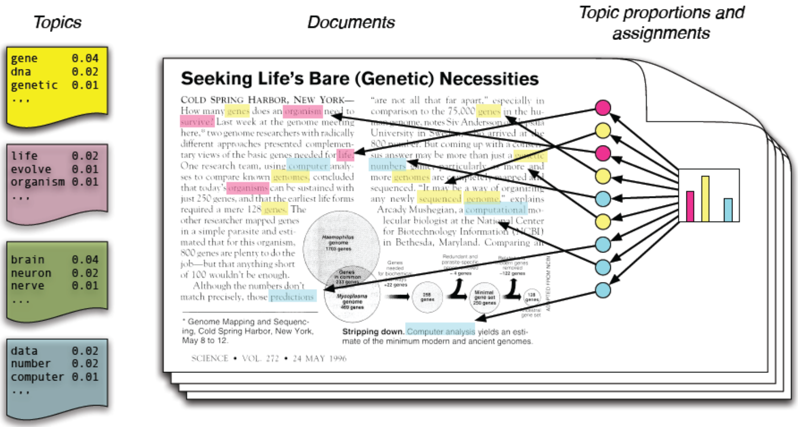
Applications (Asia pgm)



Applications (Asia pgm)

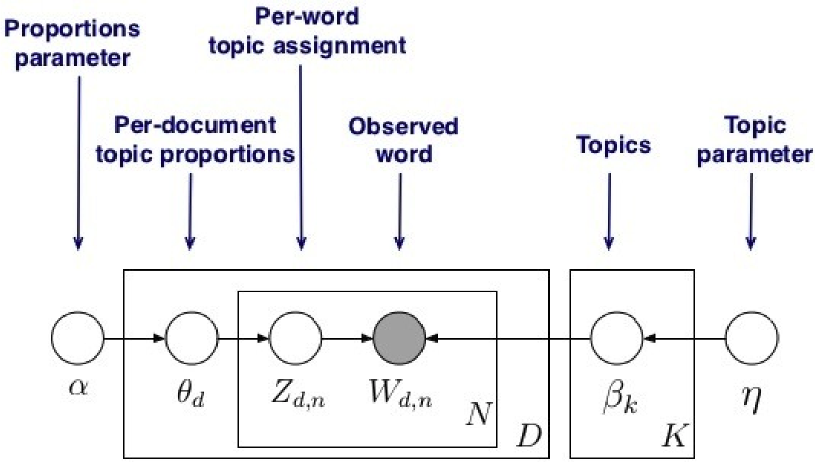


Applications *(Blei, 2012, doi: 10.1145/2133806.2133826)*



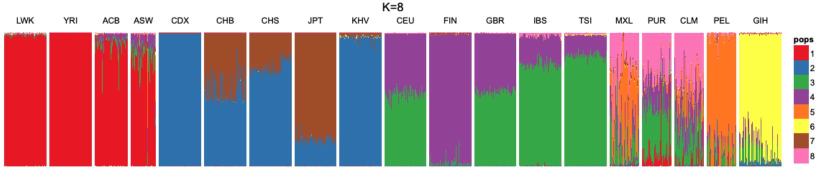
Latent Dirichlet Allocation (LDA)

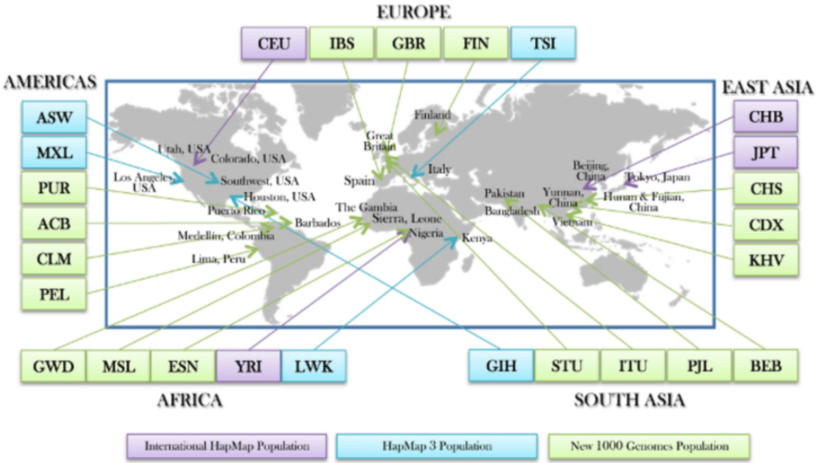
Applications (Blei, 2012, doi: 10.1145/2133806.2133826)



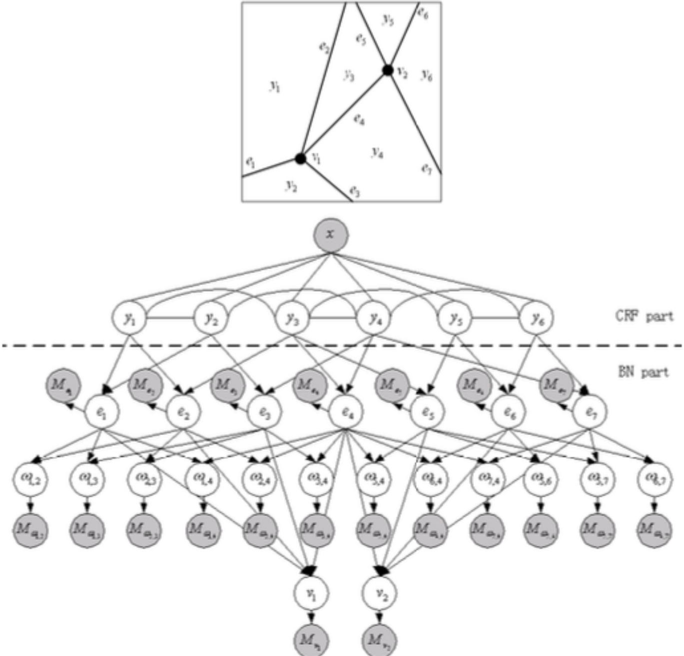
Latent Dirichlet Allocation (LDA)

Applications (Gopalan et al. 2016, doi: 10.1038/ng.3710)



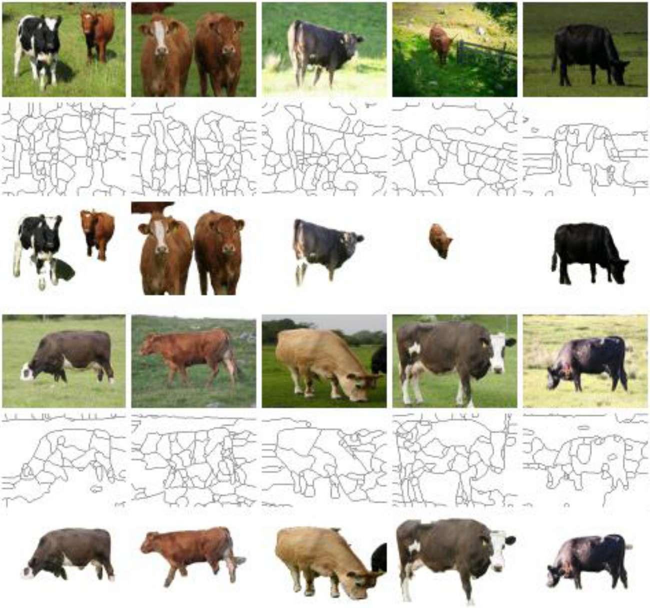


Applications (Zhang et al. 2009, doi: 10.1142/9789814273398_0006)



Applications *(Zhang et al. 2009, doi: 10.1142/9789814273398_0006)*

Applications (Zhang et al. 2009, doi: 10.1142/9789814273398_0006)



Introduction

Probabilistic Graphical Models

Jerónimo Hernández-González