Probabilistic Graphical Models

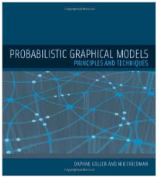
Alfredo Rossi Saldanha Cunha
Peng Yaohao

Laboratório de Aprendizado de Máquina em Finanças e Organizações

Master of Puppets

https://metacademy.org/roadmaps/cjrd/level-up-your-ml

Level 3: Master



Probabilistic Graphical Models: Principles and Techniques

There's a number of subjects you may want to study in depth at the master level: convex optimization, [measure-theoretic] probability theory, discrete optimization, linear algebra, differential geometry, or maybe computational neurology. But if you're at this level, you probably have a good sense of what areas you'd like to improve, so I'll stick with the single book recommendation. Probabilistic Graphical Models: Principles and Techniques is a classic, monstrous tomb that should be within arms length of any ML researcher worth his/her salt =). PGMs pervade machine learning, and with a strong understanding of this content, you'll be able to dive into most machine learning specialties without too much pain.

Expectations: You'll be able to construct probabilistic models for novel problems, determine a reasonable inference technique, and evaluate your methodology. You'll also have a much deeper understanding of how various models relate, e.g. how deep belief networks can be viewed as factor graphs.

Definição

Modelo gráfico probabilístico (PGM)

É um modelo probabilístico no qual um **grafo** indica a dependência condicional entre variáveis aleatórias

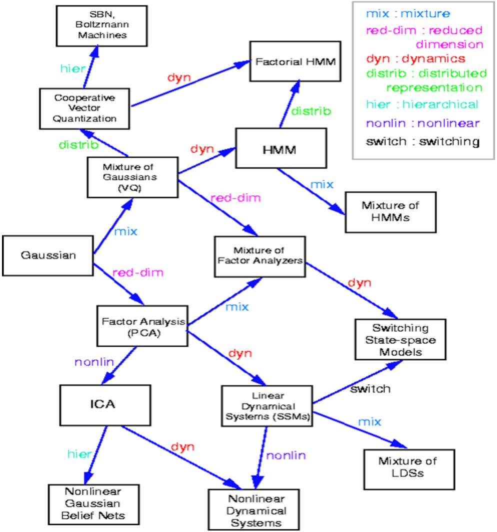
- **Grafo**: Estrutura que representa relações entre objetos com vértices e arestas

PGMs: Visão geral

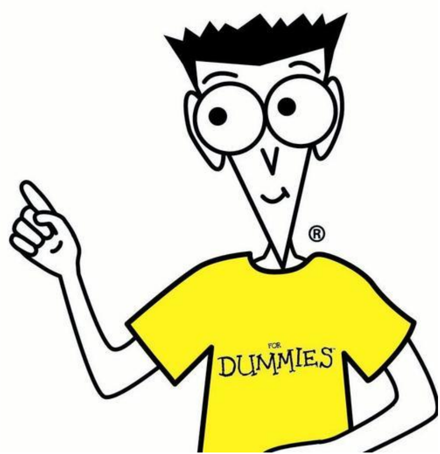
PGMs englobam uma grande variedade de modelos de estatística e *machine learning*, tais como:

- Cadeias de Markov
- Markov oculto
- Análise fatorial
- Máquinas de Boltzmann
- Redes neurais artificiais

PGMs: Visão geral



Crash Course de teoria dos grafos em 5 minutos...



https://people.revoledu.com/kardi/tutorial/GraphTheory/vertex.html

O que é um grafo?

Grafos são estruturas que "contam uma história de pontos e linhas":

- Vértices ("nós"): **coisas**
- Arestas: **relações entre as coisas**

Um grafo pode representar:

- Teclado do computador
- Árvore genealógica
- Mapa-mundi
- Fluxos migratórios
- Estrutura atômica
- ...

Grafos sem arestas são chamados de "grafos vazios"

Conexões entre nós

Nós conectados por arestas são ditos *adjacentes* (“vizinhos”)

- Uma aresta não pode conectar mais de dois nós
- Uma aresta que conecta um nó a ele mesmo é um *loop*
- Dois nós podem ser conectados por mais de uma aresta

Um grafo sem *loops* e sem arestas múltiplas é dito ser um ***grafo simples***

Grau de um nó

O grau ("valência") de um nó é o número de arestas que passa por ele

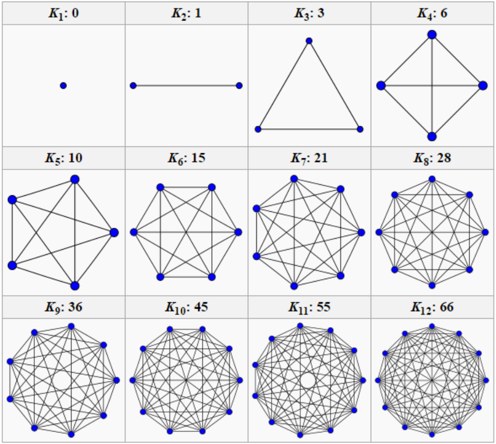
- Cada aresta contribui com 2 ao(s) grau(s) de seus destinos
- A soma dos graus de todos os nós = 2 * número de arestas ["Lema do aperto de mãos"]
- Um grafo **regular** possui todos os nós com o mesmo grau

Exercício

- Quantos nós possui um grafo regular de grau 3 que tem 6 arestas?
- É possível construir um grafo regular de grau 2 com um número arbitrariamente grande de arestas?

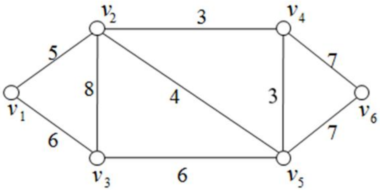
Grafos completos

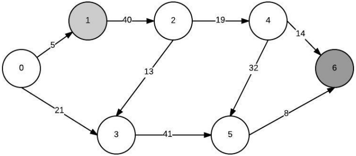
Um grafo é dito ser **completo** se todos os pares de nós forem conectados por uma aresta



Grafos direcionados e ponderados

- Grafos direcionados indicam a ordem da relação entre os nós
- Grafos ponderados possuem um valor associado a cada aresta

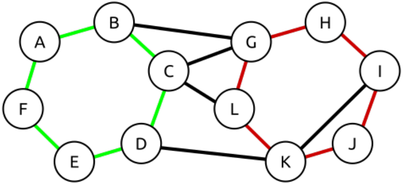


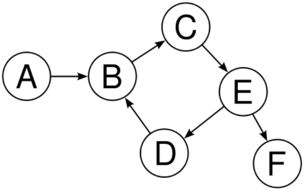


Grafos cíclicos

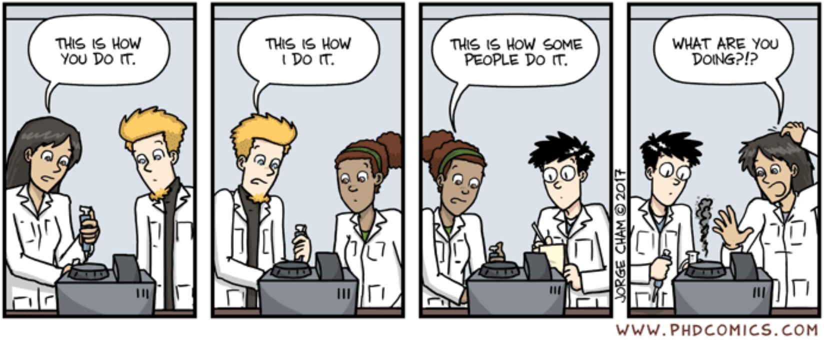
Ciclo é um caminho que começa e termina com o mesmo nó

- Um grafo **cíclico** apresenta pelo menos um *ciclo simples* (com pelo menos 3 nós, percorrendo todos os nós exceto o inicial/final apenas uma vez)





Grafos cíclicos



Bayesian Networks

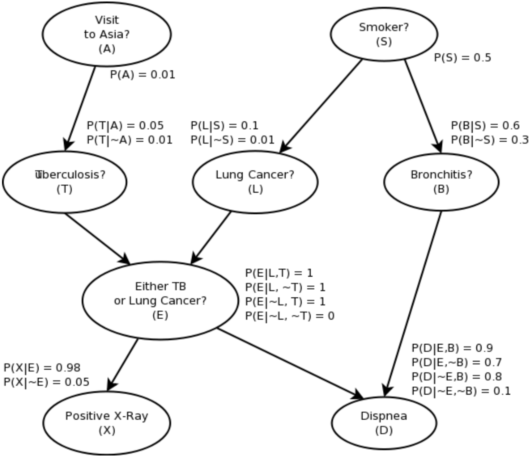
Bayesian Networks apresentam um **grafo direcionado acíclico**, no qual a distribuição conjunta das variáveis aleatórias pode ser expressa pelo produto de distribuições condicionais:

$$P[X_1, \ldots, X_n] = \prod_{i=1}^{n} P[X_i | \Xi_i]$$

Onde Ξ_i é o conjunto de nós-parente do nó X_i

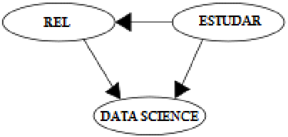
- Condicionados aos valores de seus respectivos nós-parentes, quaisquer dois nós do grafo são condicionalmente independentes
- Diminui o número de estimações da distribuição conjunta pelas condicionais e marginais

Modelos gráficos e causalidade



Modelos gráficos e causalidade

ESTUDAR	REL T	REL F
F	0.4	0.6
T	0.01	0.99



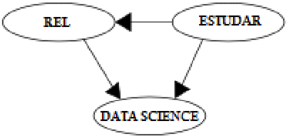
ESTUDAR T	ESTUDAR F
0.2	0.8

REL	ESTUDAR	DATA SCIENCE T	DATA SCIENCE F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- Probabilidade conjunta:
  $P(D, R, E) = P(D|R, E)P(R|E)P(E)$
- Pergunta observacional: $P(E|D) = \frac{P(D|E)P(E)}{P(D)}$

Modelos gráficos e causalidade

ESTUDAR	REL T	REL F
F	0.4	0.6
T	0.01	0.99



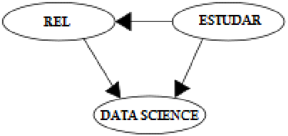
ESTUDAR T	ESTUDAR F
0.2	0.8

REL	ESTUDAR	DATA SCIENCE T	DATA SCIENCE F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- Pergunta de ação: $P(R, E|do(D)) = P(R|E)P(E)$
- $P(E, D|do(R = T)) = P(D|E, R = T)P(E)$

Modelos gráficos e causalidade

ESTUDAR	REL T	REL F
F	0.4	0.6
T	0.01	0.99



ESTUDAR T	ESTUDAR F
0.2	0.8

REL	ESTUDAR	DATA SCIENCE T	DATA SCIENCE F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- Contrafactual: $P(D|R_{\oplus\substack{\Longleftrightarrow\\\Longleftrightarrow}})$

Markov Networks

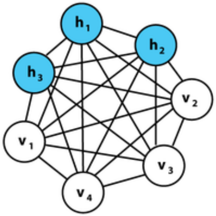
Markov Networks apresentam um **grafo não-direcionado (cíclico ou não)**, no qual:

- Quaisquer duas variáveis aleatórias são condicionalmente independentes, dadas todas as outras variáveis: $X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}$
- Uma variável aleatória é condicionalmente independente a todas as outras dada sua vizinhança fechada: $X_v \perp\!\!\!\perp X_{V \setminus \mathrm{N}[v]} \mid X_{\mathrm{N}(v)}$
- Quaisquer dois conjuntos de variáveis aleatórias são condicionalmente independentes dado um conjunto separador: $X_A \perp\!\!\!\perp X_B \mid X_S$

Máquinas de Boltzmann

Modelo não-supervisionado com grafo completo que atribui probabilidades a estados entre unidades visíveis e unidades ocultas

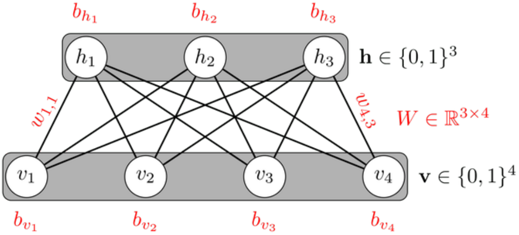
- A distribuição de Boltzmann modela a probabilidade de ocorrência de um estado num sistema físico em função da energia e da temperatura



- Extrai características mais relevantes de um fenômeno ("estados mais estáveis")
- Inviáveis de treinar na prática

Máquinas de Boltzmann Restritas

São máquinas de Boltzmann cuja representação é um grafo bipartido

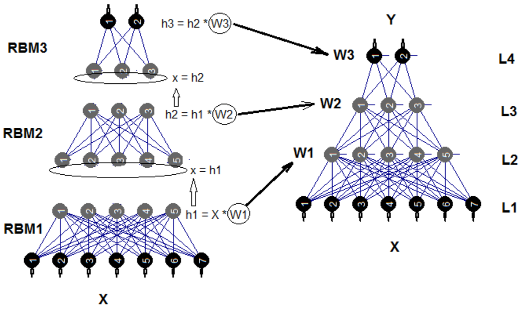


$$E(\boldsymbol{h}, \boldsymbol{x}) = -\boldsymbol{b} \cdot \boldsymbol{v} - \boldsymbol{c} \cdot \boldsymbol{h} - \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}$$

- Treinamento rápido
- Sistemas de recomendação/filtro colaborativo [Prêmio Netflix]
- Constructos latentes formativos?

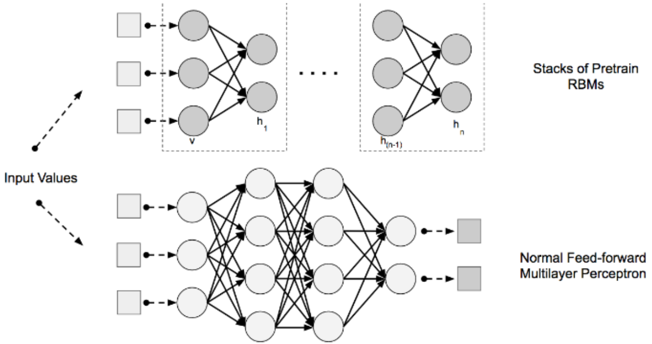
Deep Belief Network

É um empilhamento de RBMs



- Podem ser treinadas de forma *greedy*
- Procedimento eficiente, de camada por camada, para aprender os pesos de cima para baixo, sem necessidade de ter um label

Deep Belief Network



- Reconhecimento de face
- Reconhecimento de vídeo (sequência)
- Imagem de Satélite, NASA

Inferência, Estimação por amostras

- Rejection Sampling
- Importance Sampling
- **Markov Chain Monte Carlo** (MCMC)
  - Metropolis Hastings (MH)
  - Gibbs Sampling

MCMC

- Classe de algoritmos
- Muito utilizada desde a década de 80/90
- Soluciona problema mais complexos de estimação
- Consiste em propor uma distribuição (q(x)) e tirar amostras (q(x_t/x_{t-1}))
- Vai "descobrindo"a distribuição aos poucos

MH

- Constroi uma Cadeia de Markov ergótica e estacionária em relação a distribuição original
- Distribuição proposta converge quase certamente para a original
- Algortimo MH
  - $x_t \rightarrow q(x_t/x_{t-1})$, q é a dist. proposta
  - Calcule a probabilidade de aceitação $p(x_t, x_{t-1}) = min(1, \frac{\hat{p}(x_t)q(x_{t-1}/x_t)}{\hat{p}(x_{t-1})q(x_t/x_{t-1})})$
  - Aceite x_t com probabilidade $p(x_t, x_{t-1})$ e x_{t-1} com probabilidade $1 - p(x_t, x_{t-1})$

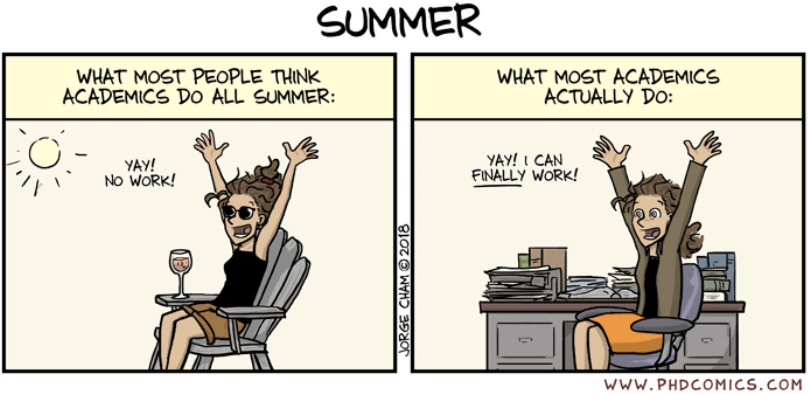
Gibbs

- Precisa das distribuições condicionais
- Gerar amostras da posteriori varrendo cada variável para amostrar a partir de sua distribuição condicional com as variáveis restantes fixadas
- Algoritmo Gibbs
  - Inicie o valor da priori $x^{(0)} \leftarrow q(x)$
  - Itere para $i = 1, \cdots$

    $x_1^{(i)} \leftarrow p(X_1 = x_1 / X_2 = x_2^{(i-1)}, \cdots, X_D = x_D^{(i-1)})$
    $x_2^{(i)} \leftarrow p(X_2 = x_2 / X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \cdots, X_D = x_D^{(i-1)})$
    $\vdots$
    $x_D^{(i)} \leftarrow p(X_D = x_D / X_1 = x_1^{(i)}, \cdots, X_{D-1} = x_{D-1}^{(i)})$

Aplicação no R

Obrigado!


SUMMER
WHAT MOST PEOPLE THINK ACADEMICS DO ALL SUMMER:
YAY! NO WORK!
JORGE CHAM © 2018
WHAT MOST ACADEMICS ACTUALLY DO:
YAY! I CAN FINALLY WORK!
WWW.PHDCOMICS.COM