Universidade de São Paulo,
Instituto de Física de São Carlos,
Grupo de Física Computacional e Instrumentação Aplicada.

Minicurso de Mineração de Dados

Condutor: Renato Fabbri
renato.fabbri AT gmail DOT com

Seminário sobre Conhecimento Aberto na Universidade
Plades / PPGDSTU / NEAE / UFPA
Prointer / ILC / UNAMA
EMBRAPA, Coletivo Casa Preta, Colaborativa

Roteiro

- Apresentação (1)
  - O que é mineração de dados
  - Utilidade
  - Histórico
  - Literatura
- Descrição alto nível (2)
  - Termos pertinentes
  - Etapas fundamentais
  - Tipos de aprendizado de máquina
  - Algoritmos paradigmáticos
- Conclusões

1- O que é?

- Mineração de dados:
  - “Processo de procurar por dados, informação ou padrões escondidos em um grupo de dados.“
  - “Data mining uses artificial intelligence techniques, neural networks, and advanced statistical tools (such as cluster analysis) to reveal trends, patterns, and relationships, which might otherwise have remained undetected.”
  - “Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information.”
  - “A process used by companies to turn raw data into useful information.”
  - “Data mining is the search for information in a seemingly endless mountain of data.“

1- Utilidade

- Usada para:
  - Association - looking for patterns where one event is connected to another event.
  - Sequence or path analysis - looking for patterns where one event leads to another later event.
  - Classification – atribuição de rótulos aos dados.
  - Clustering - finding and visually documenting groups.
  - Forecasting - discovering patterns in data that can lead to reasonable predictions about the future (This area of data mining is known as predictive analytics.).
    - Ou:
  - Predição.
  - Descrição.
- Aproveitada em:
  - cybernetics, genetics, marketing, web mining, medicine, etc.

1- Histórico

- (((Pré-termo MD:
  - Objetos de contagem – 20k AC
  - Pinturas nas cavernas – 15k AC
  - Escrita – 3,5k BC
  - Censos Babilônicos – 1,8k AC
  - Biblioteca de Alexandria – 325 AC
  - Johannes Gutenberg – 1,5k DC
  - Dicionário fonético – 1,6k DC
  - Handbook de química orgânica – 1,8k DC
  - Linguística Computacional – 1,95k DC)))
- Década de 60, armazenamento de dados em discos, fitas e computadores.
- Década de 80, bases de dados relacionais e linguagens de consultas estruturadas.
- Década de 90, primeiras aplicações relevantes:
  - OCR – Bell Labs; Rolling mills – Siemens; Séries temporais financeiras – Prediction Company.
  - Estabelecimento da área como funcional e relevante.
  - 89: First IJCAI workshop on Knowledge Discovery in Databases.
- Primeira década do milênio, impacto:
  - Visão computacional; Bio-informática; IA; Processamento de texto.
  - Novas tecnologias: SVM, SLT, Métodos graficos e bayesianos.
  - Espaço criado: Microsoft; Google; Posições acadêmicas
- Últimos 5 anos:
  - MD é termo popular

1- Literatura

- Workshops e revistas acadêmicas:
  - ACM SIG (1989 -)
  - SIGKDD Explorations
  - Lista em: http://en.wikipedia.org/wiki/Data_mining#Research_and_evolution
- Livros sobre MD:
  - Data Mining: Practical Machine Learning Tools and Techniques
- Livros sobre AM.
- Livros sobre reconhecimento de padrões.

2- Termos pertinentes

FONTES: Nomes em si e Wikipédia

- Mineração de dados: “an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.”
- Aprendizado de máquina: “a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”
- Inteligência artificial: “is the intelligence of machines or software, and is also a branch of computer science that studies and develops intelligent machines and software.”
- Reconhecimento de padrões: “In machine learning, pattern recognition is the assignment of a label to a given input value.”
- “KDD (Knowledge Discovery in Databases)”: “The non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.”

- “Big data”: “a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.” (datawarehouse, datamart)

2- Termos pertinentes (b)

Uma escolha de definições particularmente coerente:

- **Aprendizado de máquina**: tecnologia
- **Mineração de dados**: aplicação.
- "**KDD (Knowledge Discovery in Databases)**": área de aplicação da qual MD é subárea.

KDD e MD podem ser, de fato, considerados sinônimos.

2- Etapas fundamentais

- Seleção / aquisição dos dados
- Pré-processamento:
  - Remoção de ruído e elementos ruidosos
  - Normalização
  - Transformação (e.g. rotação, decomposição em alguma base, redução de dimensionalidade)
- Mineração dos dados (predição ou descrição):
  - Detecção de anomalia
  - Aprendizado de regra de associação
  - Clusterização
  - Classificação
  - Regressão
  - Sumarização (incluindo visualização e relatório)
  - Mineração de padrões sequenciais
    - OU
  - Reconhecimento de padrões
  - Clusterização
  - Classificação
- Validação / Interpretação / Avaliação

2- Etapas fundamentais (b)

- CRISP-DM: “Cross Industry Standard Process for Data Mining”
  - **Business Understanding**: “a preliminary plan designed to achieve the objectives”
  - **Data Understanding**: “first insights into the data”
  - **Data Preparation**: “ construct the final dataset from the initial raw data”
  - **Modeling**: “modeling techniques are selected and applied”
  - **Evaluation**: “a decision on the use of the data mining results should be reached”
  - **Deployment**: “the knowledge gained will need to be organized and presented”

2- Etapas fundamentais (c)

- SEMMA: “Sample, Explore, Modify, Model and Assess”
  - **Sample**: “selecting the data set for modeling”
  - **Explore**: “understanding of the data”
  - **Modify**: “select, create and transform variables”
  - **Model**: “applying various modeling techniques”
  - **Assess**: “evaluation of the modeling results”

2- Pre-processamento

- Procedimentos especialmente úteis e comuns:
  - Normalização (-media / dp)
  - Redução de dimensionalidade (PCA)
  - Decomposição em frequência

2- Tipos de Aprendizado de Máquina

- Supervisionado: inferência de uma função através de dados rotulados.
  - Análogo ao aprendizado de conceitos por humanos e animais.
  - Tipo mais comum de AM.
- Não supervisionado: evidenciamento de propriedades estruturais de dados não rotulados.
  - Análogo à observação de similaridades ou padrões por humanos e animais.
- Semi-supervisionado: uso de dados rotulados e não rotulados para a fase de treinamento.
  - Geralmente visto como híbrido entre o aprendizado supervisionado e o não-supervisionado.
  - Tem tido interesse crescente por adequação à abundância de dados.
  - Tradicionalmente entendido como próprio para os caso em que há poucos dados rotulados.
  - Tem sido melhor compreendido como pertinente mesmo nos casos com abundância de dados rotulados.

2- Algoritmos paradigmáticos (a)

- Supervisionado:
  - Redes neurais
  - Bayesiano
  - AG
  - Vizinhos mais próximos
- Não supervisionado:
  - K-means
  - Kohonen
  - ACO (Ant Colony Optimization)
- Semi-supervisionado:
  - Propagação de rótulo

2- Algoritmos paradigmáticos (b)

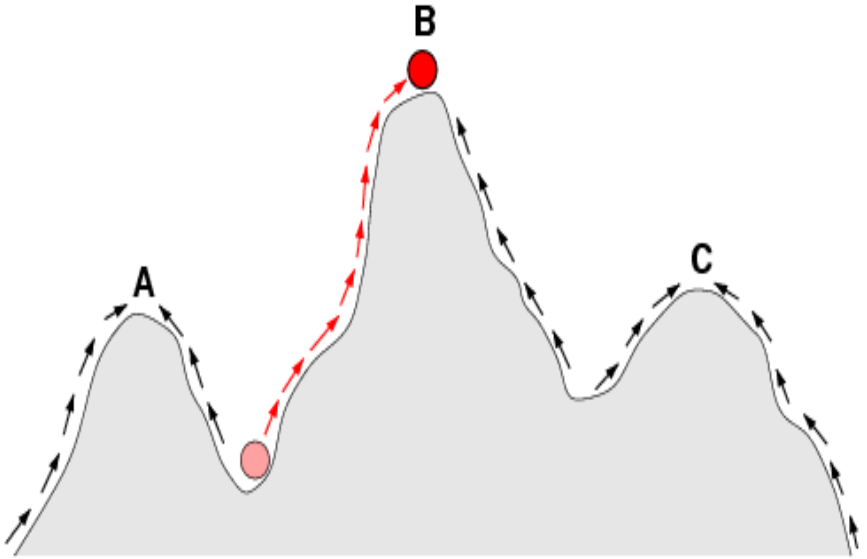
- **Hill Climbing**: If the change produces a better solution, an incremental change is made to the new solution, repeating until no further improvements can be found.

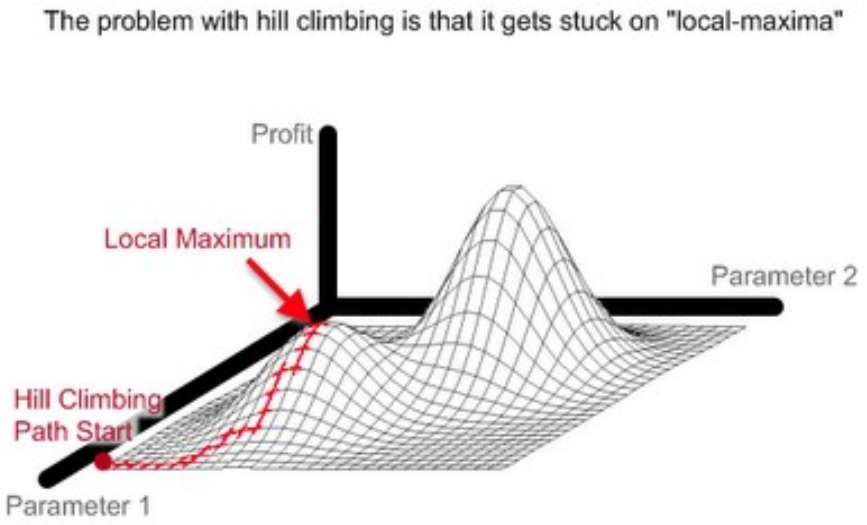
Vs

- **Simulated Annealing**: Simulated annealing (SA) is a generic probabilistic metaheuristic for the global optimization problem of locating a good approximation to the global optimum of a given function in a large search space.

2- Algoritmos paradigmáticos (b1)

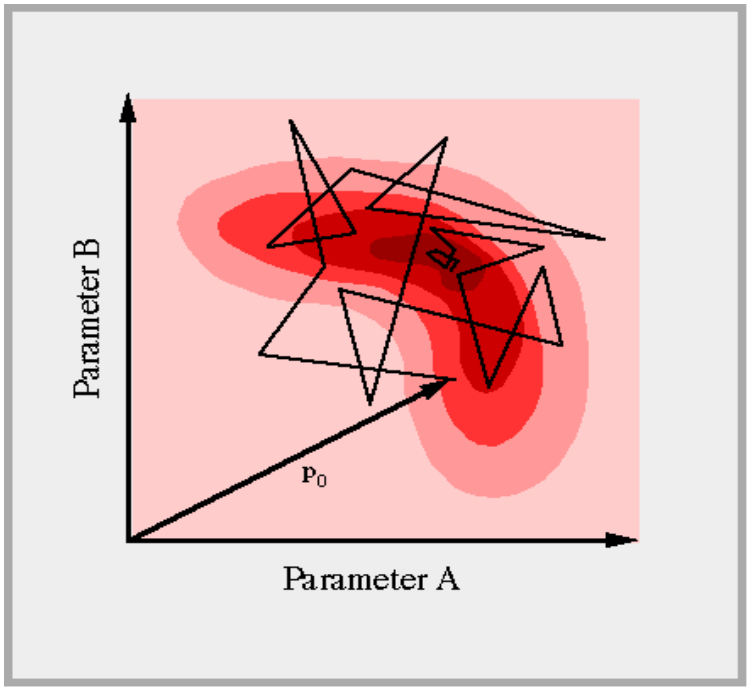
- **Hill Climbing**:
  - Pseudocódigos em: http://en.wikipedia.org/wiki/Hill_climbing
  - Implementações em python em: http://trac.assembla.com/audioexperiments/browser/NinjaML/python

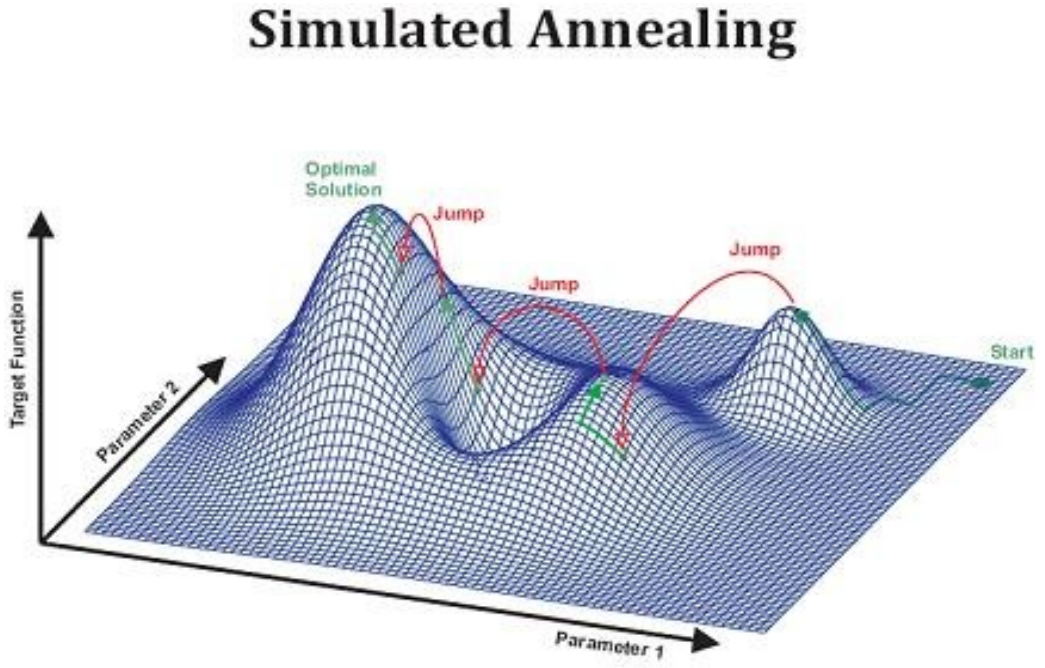




2- Algoritmos paradigmáticos (b2)

- **Simulated Annealing**:
  - Pseudocódigos em: http://en.wikipedia.org/wiki/Simulated_annealing
  - Implementações em python em: http://trac.assembla.com/audioexperiments/browser/NinjaML/python





2- Algoritmos paradigmáticos (c)

- Supervisionado:
  - Redes neurais: descrição do perceptron simples e redes neurais mais usuais.
  - Bayesiano: estrito e ingênuo (“naïve bayes”).
  - Vizinhos mais próximos.
  - Algoritmo genético.

2- Algoritmos paradigmáticos (c1)

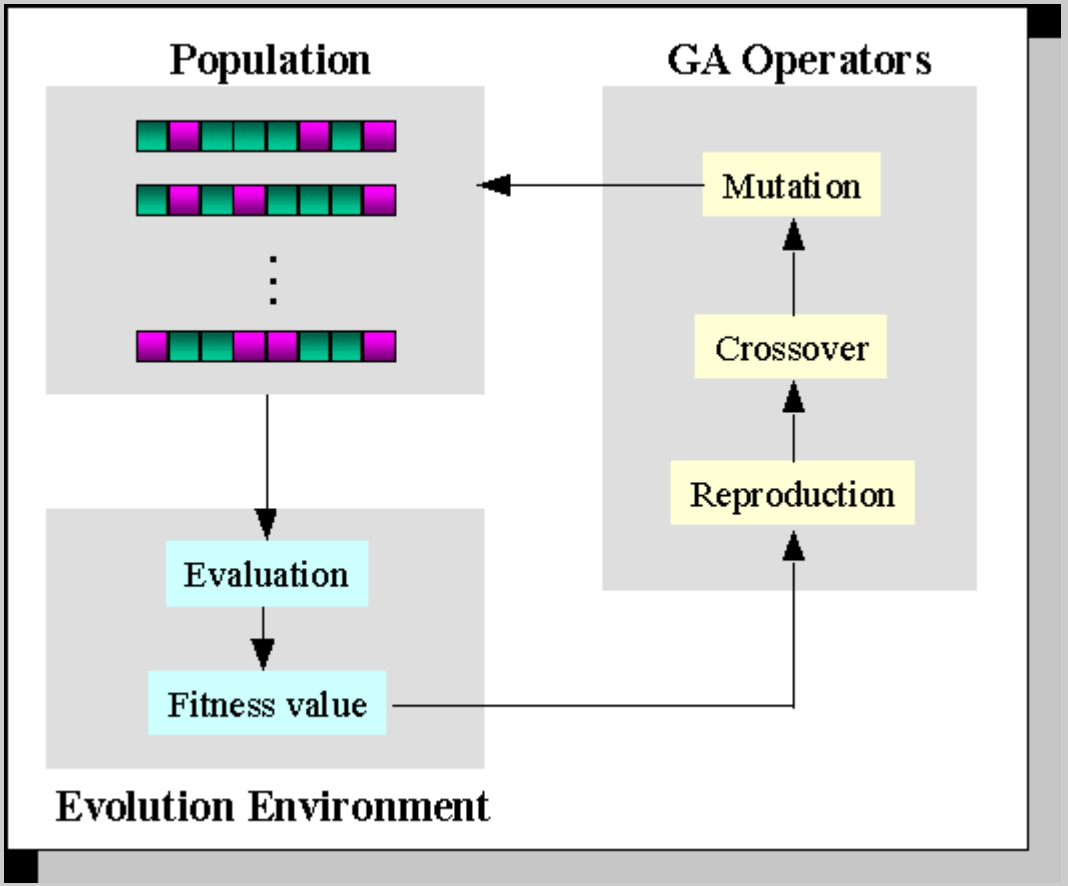
- Bayesiano: estrito e ingênuo (“naïve bayes”)
  - Qual a chance de que um elemento seja de uma determinada classe dadas suas características?
  - A suposição de independência entre as variáveis prejudica o resultado mas facilita a obtenção das probabilidades necessárias. Resolve o problema dos dados esparços.
  - Pode-se usar as condicionais com e sem independência, com pesos que contemplem os dados em mãos.
  - Descrição boa em:
    - https://en.wikipedia.org/wiki/Naive_Bayes_classifier

2- Algoritmos paradigmáticos (c1)

- Vizinhos mais próximos:
  - Atribui-se ao elemento a classe predominante nos N vizinhos mais próximos.
    - N arbitrário e geralmente ímpar.
    - Distância euclidiana
  - Descrição excelente em:
    - https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

2- Algoritmos paradigmáticos (c2)

- Algoritmo Genético:



2- Algoritmos paradigmáticos (d)

- Não-Supervisionado:
  - **K-means**:
    - minimiza a distância intra-classe
    - Descrição boa em:
      - http://en.wikipedia.org/wiki/K-means_clustering
  - ACO: Otimização por colônia de formigas

2- Algoritmos paradigmáticos (d2)

- ACO: Otimização por colônia de formigas
  - “Formigas” deixam feromonios atratores por onde as soluções são melhores.
  - Tradicionalmente usado para resolver o problema do caixeiro viajante.
  - Diz-se que a distribuição de energia elétrica é atualmente resolvida por ACO.
  - Descrição decente em:
    - http://en.wikipedia.org/wiki/Ant_colony_optimization_algorithms
  - Implementações em Python em:
    - http://trac.assembla.com/audioexperiments/browser/NinjaML/python

2- Algoritmos paradigmáticos (e)

- Semi-Supervisionado:
  - Realiza classificação considerando tanto:
    - As características dos dados rotulados
      - Quanto
    - A topologia dos dados considerados
  - Exemplos paradigmáticos:
    - Propagação de rótulo
    - Mincut
  - Explicação cuidadosa nos PDFs sobre SSL em grafos (abrir).
  - Implementações em:
    - http://trac.assembla.com/audioexperiments/browser/NinjaML

2.99- Atividades?

- Implementar algoritmos dentre os que foram apontados ou ainda outros.
- Achar algoritmos relacionados ou aprofundar entendimentos.
- Baixar o Weka ou o pyml ou o pybrain e ver os recursos disponivels.
- Buscar bancos de dados utilizáveis e potenciais bancos de dados massivos.
- Pensar em formas de obtenção de dados, por exemplo através de raspagem ou cruzamentos.
- Elaborar formas de utilização da natureza para obtenção de soluções. Por exemplo, uso de um colônia de formigas real para obter soluções do caixeiro viajante ou de distribuição de energia.
- Discussões de usos civis destas tecnologias. (Saúde, distribuição de bens e midia, etc)
- Correlação com redes sociais e dados delas provenientes.
- Amadurecimento das definições dadas e pertinência dos limiares propostos.
- Utilização dos dados dos presentes na sala. Há possibilidade de levantar um “big data”?
- Concepção de usos locais/regionais de Belém ou nos limites da atuação do NAEA ou UFPA.
- Concepção de alguma aplicação que mude o rumo das coisas ou tenha um impacto dificil de desconsiderar.

3- Conclusões

- MD muito recente: ~25 anos que despontou mas somente 10-15 anos que tomou a forma atual.
- Altamente relacionada com as áreas de aprendizado de máquina, reconhecimento de padrões e inteligência artificial. Chegando a ser pouco ou nada diferenciada por vários pesquisadores e pela literatura.
- Fundamentada em desenvolvimentos recentes, como a ampla disponibilidade de dados e recursos computacionais, avanços em métodos estatísticos e de aprendizado de máquina.
- De utilidade quasi-ubíqua, a MD é utilizada diretamente ou indiretamente por vários grupos de pesquisa. Mesmo assim, a ênfase da literatura é em usos industriais e comerciais.
- A notável disponibilidade de recursos tecnológicos abertos (e.g. Weca e bibliotecas em Python como pyml e pybrain) são facilitadores de processos científicos e mercadológicos e são entregas para potenciais usos civís.
- A MD possue traços simbióticos com áreas de conhecimento em evidência, como Redes Complexas e Processamento de Linguagem Natural.

) (/\/\/\/\ _o_o_ oOo _o_o_ /\/\/\/\ ) (

- Obrigado!!
  - Especialmente:
    - Larissa Carreira e Jader Gama.
    - Prof. Osvaldo Novais de Oliveira Jr.
    - NAEA / UFPA
    - LabMacambira.sf.net
- Email:
  - Fabbri ARROBA usp PONTO br
- Comentários, sugestões?
- Visite-nos!