Técnicas de Machine Learning I

* Definição – Uso dos dados em conjunto com os algoritmos para a produção de informação que serão relevantes para a tomada de decisão.
  + Informações uteis (Data-driven decision making)
  + Exemplo: criação de modelos preditivos e/ou analisando a interdependência entre os dados

Algoritmos com implementações de métodos estatísticos.

* Banco de dados
  + Definição: é o objeto onde estão armazenadas as informações de interesse para a análise ou estudo em questão. Onde estão as informações históricas.
  + Em muitos casos o Baco de dados contém uma amostra, ou seja, é um subconjunto extraído da população.
  + O banco de dados é composto por variáveis e observações.
    - Observações: as unidades de análise que têm suas características e atributos medidos
    - Variáveis: características, atributos observados, medidos ou categorizados. Atributos de interesses. Características que estamos medindo
  + Exemplos
    - Pessoas
    - Países
    - Empresas
    - Tarefas
    - Ações da bolsa

O que define a unidade de observação é basicamente o seu objetivo de análise.

* Estrutura para uso

Normalmente usamos uma estrutura tabular com as variáveis em colunas e observações em linhas.

Insumo básico são os dados.

Técnicas estatísticas estão sempre por traz.

* Tipos de Variáveis
  + Métricas: são as variáveis quantitativas, isto é, apresentam características que podem ser mensuradas ou contadas muitas vezes. São chamadas de variáveis quantitativas.
  + Não Métricas: São as variáveis qualitativas, sendo que indicam características que não podem ser medidas. Tais variáveis contem categorias, por isto, muitas vezes são chamadas de variáveis categóricas.
* A identificação do tipo de variável é fundamental para a escolha da técnica que será utilizada na análise dos dados. Ou seja, da escolha de qual método estatístico que pode ser aplicado à variável.

Exemplos:

* + Métricas
    - Idade em Anos
    - Renda mensal em reais
    - Distância entre duas cidades em metros
    - Rentabilidade percentual diária de uma ação
    - Número de habitantes em um município
  + Não Métricas
    - Nacionalidade
    - Cor do Veículo
    - Profissão
    - Grau de escolaridade
    - Respostas sim ou não a um questionário
    - Escalas Likert
* Escala Likert: Respostas de questionários em formato de escala que permite mensurar o ponto de vista e a postura do entrevistado indo por exemplo de discordo totalmente a concordo totalmente.
* Variáveis métricas são sempre numéricas
* Variáveis qualitativas normalmente são textuais, porém podem vir disfarçadas de números.
  + Ex: 1 - Muito bom, 2 - Médio, 3 - Ruim

Apesar dos números ainda é uma variável qualitativa

Cada tipo de variável tem técnicas e métodos estatísticos apropriados

Exemplos:

* Variáveis qualitativas:

Tem a sua representação em termos descritivos basicamente por meio de tabelas de frequências e gráficos derivados delas. Ou seja, para elas, não é possível aplicar medidas de resumo como média ou desvio padrão. Estatísticas como média e desvio padrão são estatísticas aplicadas exclusivamente para variáveis métricas.

* Tabela de Frequência: Realiza uma contagem de frequências observadas, contagem por categoria.
* Não é recomendado atribuir labels numéricos para variáveis métricas, pois gera ruido na analise.
* Variáveis quantitativas

As variáveis quantitativas podem ser representadas por diversas ferramentas, como gráficos, medidas de posição e dispersão.

Exemplos de estatísticas descritivas

* + Medias de posição: Média, Mediana, Quartis
  + Médias de dispersão: Variância e desvio padrão
* Outras características relevantes
  + Variáveis qualitativas: dicotômico ou policotômica, Nominal ou ordinal
    - Dicotômica: duas categorias (binaria)
    - Policotômica: mais de duas categorias
    - Nominal: não estabelece relação de grandeza\ordem
    - Ordinal: Estabelece Ordem
  + Variáveis Quantitativas: Discretas ou contínuas.
    - Discretas: Possuem conjunto Finito e numerável de valores, em geral são obtidas a partir de dados de contagem (0,1,2,3,4,5)
    - Contínuas: Assumem valores pertencente ao intervalo de números reais.
* É possível categorizar uma variável métrica, uma variável originalmente métrica pode ser categorizada sem problema.
* Porém o inverso não é verdadeiro, não é possível transformar uma variável originalmente categórica para métrica, pois ao realizar este procedimento se incorre no que se chama de um erro de ponderação arbitraria. Pois seria necessário atribuir pesos arbitrários para as categorias, um procedimento errado e muito recorrente. Não podemos fazer ponderação arbitraria.

FIM bloco 1

Bloco 2

Machine learning é um subconjunto de IA

* Data Frame = Representação de um banco de dados.
* Modelos lineares de regressão
  + Modelos supervisionados de machine learning
    - Conhecidos como modelos confirmatórios ou técnicas de dependência
    - O objetivo é estimar modelos, equações, com o intuito de elaborar previsões. Portanto, há inferência de resultados para outras observações fora da a mostra.
    - Define-se uma relação Y= F(x)
      * Y: Chamada de Variável dependente é a variável a ser explicada (target)
      * X: chamada de variáveis explicativas são as preditoras (features)
* Quando Aplicar o modelo ?
  + A regressão Linear é aplicada quando a variável dependente é quantitativa
    - O objetivo é explicar o comportamento de Y em função de um conjunto de X
    - Estabelece-se uma relação Linear entre as variáveis.
* Regressão Linear simples e múltipla
  + A regressão Linear simples contém apenas uma variável explicativa
  + A regressão Linear múltipla contém mais de uma variável explicativa.
* Modelo geral de regressão Linear

Yi = a + b1\*x1i+b2X2i+...bkXki+ui

* + Y: é a variável dependente quantitativa
  + a: Representa a constante (Intercepto)
  + :Representam os coeficientes para cada variável explicativa
  + :Representam as variáveis explicativas do modelo
  + : representa o termo de erro do modelo
    - : é o número de variáveis explicativas
    - : se refere as observações em análise
  + As variáveis explicativas (x) podem ser métricas ou categóricas.

O papel do algoritmo é estimar os parâmetros A e B que mostram o intercepto e os coeficientes para cada variável explicativa.

Estimar o modelo Significa estimar os valores para o a(alpha) e b(beta)

* Mínimos quadrados Ordinários (MQQ)
  + o algoritmo estimara os parâmetros e do modelo

Ychapel = Aplha + Beta + Xi

* + Pode se definir o resíduo do modelo para dada observação i:

Ui – Yi – Ychapel i

* + Condições pada a estimação dos parâmetros do modelo (MQO)
    1. A somatória dos resíduos deve ser igual a zero
    2. A somatória dos resíduos ao quadrado é a mínima possível
* Pode ser visto como P.O.

Fim boco 2

Bloco 3

Modelos supervisionados precisam ser definidos por uma função de Y em função de X, ou seja, o proponente é quem define as relações, por este motivo este modelo possui capacidade preditiva.

No treinamento não supervisionado, o algoritmo realiza a análise da interdependência entre as variáveis, ou seja, a análise exploratória das variáveis. No treinamento não supervisionado não há a proposição da função Y em função de x, neste caso os algoritmos buscam padrões sem que seja informado a relação previamente determinada.

No modelo supervisionado, é proposto um modelo preditivo, para que sejam feitas inferências para observações de fora da amostra.

Já no treinamento não supervisionado não é gerado o modelo preditivo. Ele é mais utilizado para o diagnóstico da amostra, modelos não supervisionados não tem capacidade preditiva para observações fora da amostra.

MQO (mínimos quadrados Ordinários

OLS ( Ordinary List Sqares)

É um método de obtenção da reta de regressçao do modelo

Pode ser visto como um problema de P.O. (Pesquisa operacional)

Como são estabelecidas as condições para a estimação dos parâmetros alpha e beta

1. Primeira condição: a somatória dos resíduos deve ser igual a 0. Ou seja, a somatória dos “erros” deve ser 0. Porém com isso podem existir varias retas que satisfaçam essa condição.
2. Por isso temos a segunda condição: a somatória dos resíduos ao quadrado, deve ser a mínima possível. Com essas duas condições chegamos a uma única reta possível.