Métodos matriciais e Cluster analysis

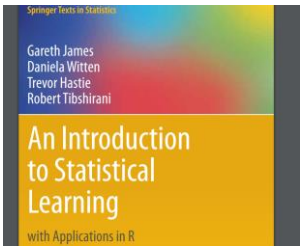
Introdução

Prof. Abraham Laredo Sicsu

Sobre as aulas

- Curso preparado para usuários de modelos preditivos.
  - Não vamos discutir como os softwares foram programados
- Instrutor exporá as principais ideias e o uso do software R.
- Livro recomendado (disponibilizado pelos autores na net):

https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf



Critério de notas cursos síncronos online (ZOOM)

- Trabalho em grupos de no MÁXIMO TRÊS participantes.....30%
  - Trabalho deve ser entregue até a data definida pelo Professor e postado no ECLASS, utilizando o ícone "entrega de atividades"
  - Nota depende de apresentação e conteúdo.
- Prova final / trabalho final 70%
  - O professor colocará um documento (.doc) em Word ("**folha de questões**") com as questões, bem com um arquivo em Excel com os dados. Vocês serão avisados da postagem por e mail.
  - Haverá diferentes conjuntos de dados. Cada aluno deve tomar cuidado para utilizar as planilhas de dados que correspondem à turma em que ele for alocado.
  - Os alunos deverão resolver as questões e colocar as respostas no documento "**folha de questões**" em Word, **sem esquecer de colocar o nome**
  - Alunos poderão consultar todos os materiais disponíveis no ECLASS de sua turma, acessar a internet e arquivos pessoais.
  - Após completar a prova, os alunos devem **preencher as respostas** na "**folha de questões**" , salvar a folha de prova como .pdf e postar a "**folha de questões**" no eclass, clicando no ícone "entrega de atividades" e na pasta com o título de **entrega de trabalho final**. Por favor, não copiem os scripts do R
  - **Os alunos terão um tempo limite para postar a prova no e class que será definido pelo Professor. Findo esse prazo, o sistema não aceita a postagem de novos documentos. E, por favor, não enviem por e mail pois não serão aceitos!**
  - **Comportamento ético**
    - Os alunos deverão resolver as questões **INDIVIDUALMENTE**
    - Proibido qualquer tipo de comunicação, qualquer que seja a mídia utilizada (oral, eletrônica, sinais de fumaça...)

Horários

- **Manhã**

8h30 às 9h45 - aula

9h45 ás 9h55 – break

9h55 às 11h15- aula

11h15 ás 11h30 – break

11h30 às 12h30 - aula

- **Tarde**

13h30 às 14h45 - aula

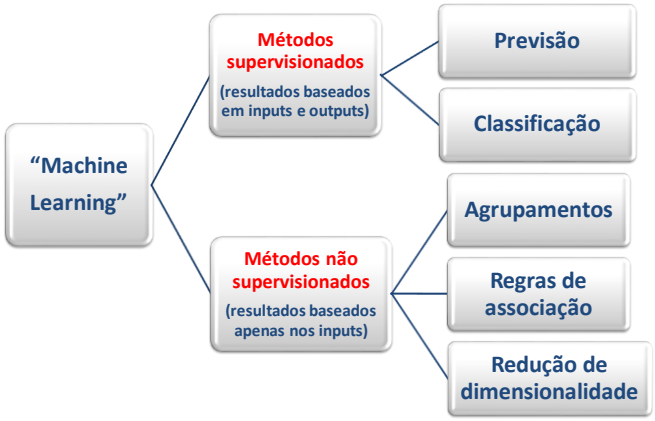
14h45 ás 14h55 – break

14h55 às 16h15- aula

16h15 ás 16h30 – break

16h30 às 17h30 – aula

Métodos



Apesar do nome, na realidade, nem todas as técnicas que serão apresentadas são "machine learning" ao pé da letra. O Analista fornece um modelo específico

Classificação dos métodos

- **Métodos supervisionados**

X_1	X_2						X_P	Y

- **Métodos não supervisionados:**

X_1	X_2						X_P

Classificação dos métodos

- **Métodos supervisionados (nossa disciplina):**
  - Objetivo: modelar a relação entre uma variável alvo (ou variável dependente) e um conjunto de variáveis previsoras.
  - Duas categorias
    - Modelos de previsão → alvo é prever o valor de uma grandeza quantitativa (demanda, valor do aluguel, vendas,...)
    - Modelos de classificação → objetivo é classificar em uma das categorias da variável alvo (bom/mau pagador, cliente de alto, médio ou baixo potencial) a partir das características definidas pelas previsoras
- **Métodos não supervisionados:**
  - Objetivo: desvendar padrões de comportamento existentes nos dados.
  - Não existe uma variável alvo para supervisionar a busca desses padrões

Métodos apresentados nesta disciplina

- Cluster analysis
- Análise das componentes principais
- Regras de associação
- Operações com matrizes
- Auto valor e auto vetor

Alguns erros usuais de analistas inexperientes

- **Não discutir o problema** com as pessoas envolvidas no processo ao qual serão aplicados os resultados e com experts no contexto do problema.
- Preocupar-se mais com a "precisão" do modelo que com sua **aplicabilidade e a interpretação dos resultados** no contexto do problema – o modelo precisará ser "vendido" dentro da empresa.
- Não alocar boa parte do tempo à **analise e interpretação de cada variável** considerada no desenvolvimento do modelo

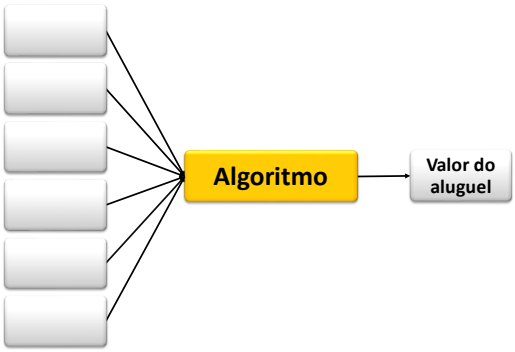
Boa leitura:

https://www.analyticsvidhya.com/blog/2018/07/13-common-mistakes-aspiring-fresher-data-scientists-make-how-to-avoid-them/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+AnalyticsVidhya+%28Analytics+Vidhya%29

Dados

Exercício: APP GVRENT

- Objetivo: Prever o preço de aluguel de um apto
- Como definir a resposta (aluguel ou “pacote”)
- Que variáveis previsoras (*inputs*) utilizar



Variáveis

Qualidade do dados

- O que é um mau pagador?
- O que é um bom funcionário?
- Como definir "experiência"?

- Importante :
  - definição operacional da variável
    - consenso
    - uniformidade de interpretação
  - processo de medição / cálculo

15

Identificação das variáveis previsoras

- Fundamental
- Ouvir experts na área de trabalho
- Brainstorming
- Na dúvida, testar?
- Importante: definição operacional

Análise dos Dados

Objetivos:

- entender o comportamento de cada variável
  - distribuição / outliers / missing values / etc..
- entender relação entre variáveis
- Fundamental para o analista ➔ "insight"

Análise Preliminar dos Dados

- Dois tipos de análise
  - Univariada
    - Analisa cada variável individualmente sem verificar relações com outras variáveis
    - Medidas descritivas , diagramas de barras, histogramas, box-plot, etc.
  - Bivariada
    - Analisa a relação entre duas variáveis do projeto
    - Em geral, foco é na relação entre previsor (X) e variável alvo (Y)
    - Correlações, medidas descritivas em cada grupo, box plot, tabelas de contingência. Diagramas de dispersão, matrizes de dispersão etc.

19

Feature engineering (preparação dos dados)

Feature engineering: tratamento da base de dados para aprimorar a análise e modelagem

1. Identificação e tratamento de MV
2. Identificação e tratamento de outliers
3. Criação de novas variáveis a partir das existentes
   - Quantificação de qualitativo
   - Transformação de variáveis
   - Etc.
4. Redução de dimensionalidade