Fundamentos da Aprendizagem de Máquina

Lista de Atividades

Aula - Conceitos sobre Aprendizagem de Máquina

Ricardo Augusto (ricardojunior@inatel.br)

Inatel

Índice

Conceitos sobre Aprendizagem de Máquina

1. Atividades

Esse arquivo consiste em uma lista de atividades a serem realizadas para o módulo sobre conceitos da aprendizagem de máquina do curso de introdução à ciência de dados e decisões. A lista é composta pelas seguintes atividades:

- **Exercícios Conceituais**
  - São dez (10) questões abertas sobre conceitos introdutórios de machine learning. O objetivo é exercitar a escrita para que ideias e conceitos de machine learning possam ser consolidados pelos próprios alunos.
- **Exercícios de Múltipla Escolha**
  - São dez (10) questões de múltipla escolha sobre os fundamentos de machine learning discutidos na aula - Conceitos de Aprendizagem de Máquina.
- **Exercícios de Revisão - Probabilidade/Estatística**
  - São duas (2) questões de revisão sobre conceitos básicos de probabilidade e estatística. De forma específica, temos uma questão relacionada à probabilidade e outra sobre o conceito de população e amostra, importante na área da estatística.
- **Exercícios Computacionais**
  - São oito (8) exercícios computacionais e alguns deles serão solucionados para consolidar os conhecimentos adquiridos em aula. Ainda assim, a solução desses exercícios deve ser entregue por cada aluno para o professor.

Cada tipo de exercício da lista de atividades terá uma nota entre 0 e 100. Assim, a composição da nota avaliativa desse módulo, denotada como N1, é dada pela combinação linear dos exercícios relacionados com as atividades citadas, considerando pesos equilibrados, de acordo com

$$\begin{aligned} \text{N1} = 0.25 \times \text{Exercícios Conceituais} + 0.25 \times \text{Exercícios de Múltipla Escolha} \\ + 0.25 \times \text{Exercícios de Revisão} + 0.25 \times \text{Exercícios Computacionais} \end{aligned} \tag{1.1}$$

Data de entrega: dia 15/05/2020 (15 de maio)

1.1 Exercícios Conceituais

1. **Com suas palavras, forneça uma definição para a aprendizagem de máquina.**

2. **Cite, pelo menos, três problemas reais nos quais técnicas de Machine Learning poderiam ser utilizadas.**

3. **Diferencie aprendizagem supervisionada da não supervisionada.**

4. **Qual o significado dos dados de treinamento rotulados (label traning dataset)?**

5. **Defina, com suas palavras, o que é um modelo de machine learning.**

6. **Que tipo de algoritmo de machine learning, em termos de categoria, poderia ser usado para segmentar clientes em múltiplos grupos?**

7. **Explique, com suas palavras, as principais diferenças entre aprendizagem *online* e *offline***

8. **Qual é a diferença entre os parâmetros e hiperparâmetros em um modelo de ML?**

9. **Explique a diferença entre os modos de aprendizagem que são baseados em modelos ou instâncias.**

10. **Se um modelo de ML atinge um bom desempenho sobre os dados de treinamento, mas não generaliza bem para novos dados (teste), o que pode estar acontecendo? O que poderia ser realizado para melhorar a generalização do modelo de ML?**

1.1.1 Respostas - Exercícios Conceituais

1.2 Exercícios de Múltipla Escolha

1. **Exercício 1 (Fundamentos de Machine Learning)**

   Considere um algoritmo de aprendizagem de máquina que interpreta marcações de e-mail (*spam* ou não *spam*) realizadas por um usuário. Baseado nesta observação, o algoritmo aprende a filtrar os e-mails de forma mais eficaz. Neste caso, a tarefa **T** da definição de aprendizagem de máquina consiste em

   - a) Classificar um e-mail como *spam* ou não *spam*.
   - b) Observar as marcações de e-mail como *spam* ou não *spam*.
   - c) O número ou razão de e-mails corretamente classificados como *spam* ou não *spam*.
   - d) Não é possível aplicar aprendizagem de máquina neste caso do enunciado.

2. **Exercício 2 (Fundamentos de Machine Learning)**

   Considere um algoritmo de aprendizagem de máquina que interpreta marcações de e-mail (*spam* ou não *spam*) realizadas por um usuário. Baseado nesta observação, o algoritmo aprende a filtrar os e-mails de forma mais eficaz. Neste caso, a métrica **P** da definição de aprendizagem de máquina consiste em

   - a) Classificar um e-mail como *spam* ou não *spam*.
   - b) Observar as marcações de e-mail como *spam* ou não *spam*.
   - c) O número ou razão de e-mails corretamente classificados como *spam* ou não *spam*.
   - d) Não é possível aplicar aprendizagem de máquina neste caso do enunciado.

3. **Exercício 3 (Fundamentos de Machine Learning)**

   Considere um algoritmo de aprendizagem de máquina que interpreta marcações de e-mail (*spam* ou não *spam*) realizadas por um usuário. Baseado nesta observação, o algoritmo aprende a filtrar os e-mails de forma mais eficaz. Neste caso, a experiência **E** da definição de aprendizagem de máquina consiste em

   - a) Classificar um e-mail como *spam* ou não *spam*.
   - b) Observar as marcações de e-mail como *spam* ou não *spam*.
   - c) O número ou razão de e-mails corretamente classificados como *spam* ou não *spam*.
   - d) Não é possível aplicar aprendizagem de máquina neste caso do enunciado.

4. **Exercício 4 (Métricas de Desempenho)**

   A avaliação de performance ou desempenho de modelos de machine learning é um ponto de relevância e, de fato, temos uma etapa de **avaliação** que pode fazer parte de um projeto de ciência de dados. Uma métrica amplamente conhecida na literatura é colocada abaixo:

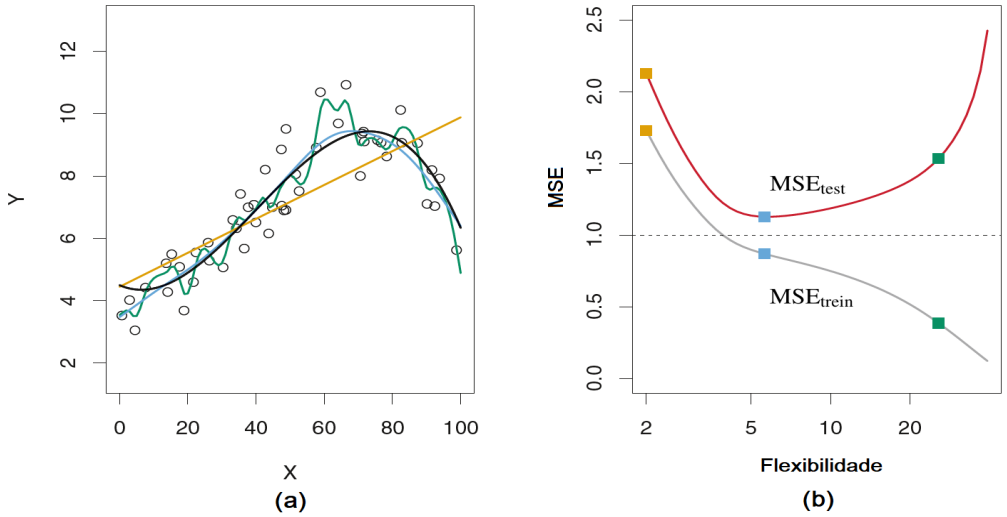
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{h}(\mathbf{x}_i))^2 \tag{1.2}$$

   Sobre a interpretação dessa métrica de desempenho, marque a alternativa correta:

   - a) A aplicação da métrica MSE só ocorre na aprendizagem não supervisionada.
   - b) Na equação, o termo y_i consiste na i-ésima estimativa do modelo de ML.
   - c) Só é possível estimar, empiricamente, o MSE para o conjunto de treinamento.
   - d) Quanto maior é o valor do MSE, pior será o desempenho do modelo de ML avaliado.

5. **Exercício 5 (Métricas de Desempenho)**

   As Figuras (a) e (b) abaixo, extraídas do livro *An Introduction to Statistical Learning*, discutem a relação entre o $\text{MSE}_{\text{trein}}$ e MSE_{test}, ou seja, o desempenho dos modelos de ML nas fases de treinamento e teste. Sobre tal relação, marque a alternativa correta:



   - a) A razão pela qual o MSE_{test} não segue o decaimento do $\text{MSE}_{\text{trein}}$ reside na falha de generalização do modelo *smoothing splines* utilizado.
   - b) O comportamento em "U" para curva do MSE_{test} ocorre porque a função hipótese verdadeira é do tipo não linear.
   - c) Nota-se que existe uma garantia de performance de teste (i.e., baixo MSE_{test}) se nós ajustarmos o modelo de ML com os dados de treinamento.
   - d) A diferença entre $\text{MSE}_{\text{trein}}$ e MSE_{test} é explicada pelo fato de que o processo de aprendizagem das técnicas de ML se baseia na minimização do $\text{MSE}_{\text{trein}}$ e, por conta disso, não pode garantir ótima generalização para os dados de teste (i.e., baixo MSE_{test}).

6. **Exercício 6 (Técnicas de ML)**

   No estudo de machine learning, realizar a associação de técnicas de aprendizagem de máquina de acordo com o supervisionamento aplicado no processo de treinamento do modelo é um aspecto importante. Sobre esse tópico, marquea alternativa correta:

   - a) O algoritmo K-Nearest Neighbors ou K-Vizinhos mais próximos pode ser aplicado em problemas somente de forma não supervisionada.
   - b) Não existe uma relação entre os algoritmos - Árvores de Decisão (Decision Threes) e Florestas Aleatórias (Random Forests).
   - c) Redes neurais artificiais podem ser usadas em problemas considerando a aprendizagem supervisionada e não supervisionada.
   - d) A análise de componentes principais (PCA) é uma técnica supervisionada de aprendizagem de máquina.

7. **Exercício 7 (Desempenho de Classificadores)**

   Um hospital conta com uma equipe de pesquisadores em ciência de dados e inteligência artificial, avaliando diversos classificadores construídos para análise de problemas pulmonares dos pacientes. O objetivo consiste em levantar a performance dos classificadores para compreender, adotar e posteriormente testar os melhores modelos treinados. Cada classificador atua para apontar riscos de doenças pulmonares em futuros pacientes, isto é, o resultado da aplicação de modelo de ML aponta a presença ou ausência de risco de doença pulmonar de um determinado paciente que dá entrada no hospital com sintomas relacionados à parte respiratória-pulmonar*.

   O exemplo abaixo consiste nos resultados de desempenho sobre o treinamento de um classificador construído a partir de centenas de imagens médicas pulmonares armazenadas no banco de dados de um hospital.

   Matriz de Resultados de Treinamento do Classificador

		Predição do Classificador	
		Risco presente	Risco ausente
Realidade	Risco presente	VP = 100	FN = 70
	Risco ausente	FP = 150	VN = 1200

   Considerando as informações apresentadas, marque a alternativa correta:

   - a) A acurácia do classificador é 85%, caracterizando o desempenho de forma completa.
   - b) A precisão calculada permite dizer que o classificador tem alto desempenho.
   - c) O desbalanceamento dos dados, especialmente com a quantidade de pacientes sem risco, aumenta a acurácia do modelo.
   - d) Falsos positivos e negativos têm impacto equivalente sobre o suporte às decisões dos resultados do classificador.

*Nesse link: http://jtd.amegroups.com/article/view/19479/html - temos um artigo científico que faz uma revisão completa sobre esse assunto, considerando o uso de redes neurais profundas (i.e., Deep Learning) para esse propósito (suporte para detecção e análise de doenças pulmonares).

8. **Exercício 8 (Técnicas de ML)**

   Alguns livros de ML trazem a informação de que a aprendizagem não supervisionada é desafiadora, no comparativo com o treinamento de modelo de forma supervisionada. Claro, o treinamento de um modelo não supervisionado ocorre de modo diferente pelo simples fato de não possuirmos dados rotulados com as saídas conhecidas. Nesse sentido, a aprendizagem não supervisionada é conduzida como parte da análise exploratória de dados e uma das técnicas mais conhecidas da aprendizagem estatística é a análise de componentes principais, PCA - Principal Component Analysis.

   Marque a alternativa que descreve corretamente qual é o conceito fundamental da PCA:

   - a) A PCA é uma técnica de aprendizagem de máquina não supervisionada e se refere a um processo de cômputo dos componentes principais de um grupo maior de variáveis (características), permitindo descrever a variabilidade estatística contida nos dados com um grupo menor de variáveis.
   - b) PCA é uma técnica de aprendizagem semisupervisionada e se refere a um processo de aglomeração de variáveis a partir de um conjunto menor de características, chamadas de componentes principais.
   - c) A PCA é uma técnica de aprendizagem não supervisionada e se refere ao processo de predição de variáveis ou características de interesse, a partir de componentes principais não ortogonais entre si.
   - d) A PCA é uma técnica de aprendizagem não supervisionada e se refere a um processo de visualização de dados para que todas as dimensões e características possam ser analisadas por meio de gráficos de correlação entre todas as variáveis.

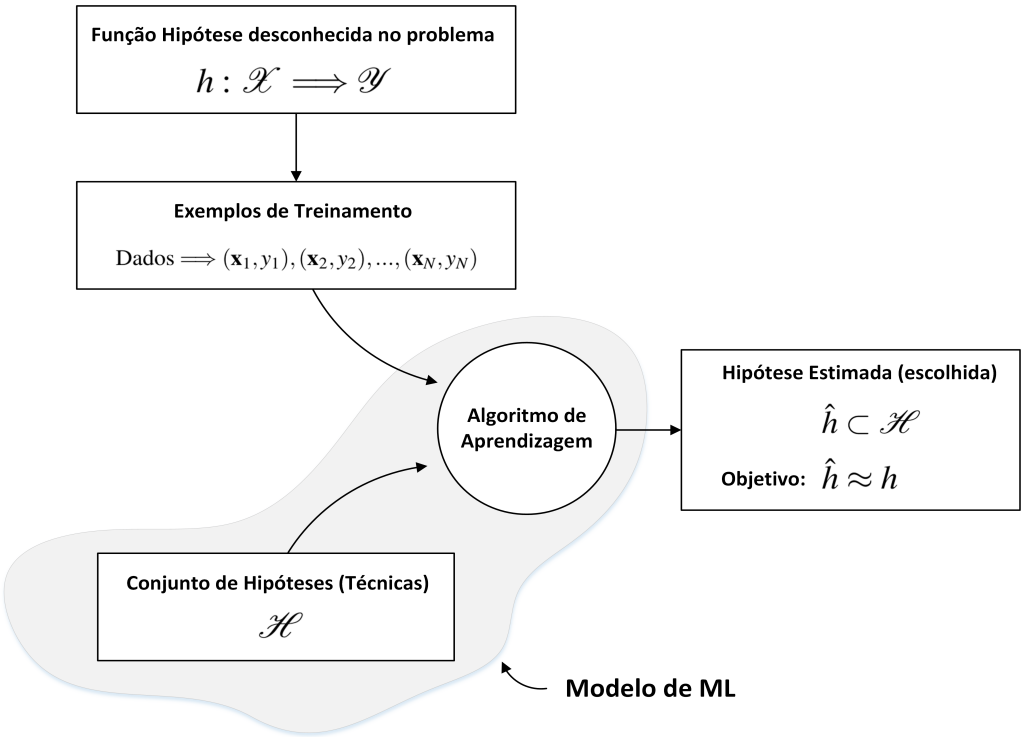
9. **Exercício 9 (Modos de Aprendizagem)**

   Na categoria modos de aprendizagem, marque a alternativa que descreve corretamente qual é a diferença de generalização entre modo de aprendizagem baseado em instâncias e modelos:

   - a) Ambos os modos de aprendizagem, baseados em instâncias e modelos, se fundamentam apenas em métricas de similaridade.
     - Na classificação de um potencial motorista para os serviços de corrida (aplicativos de transporte privado como Uber) as distâncias entre as coordenadas de localização de uma pessoa (por meio de seu smarphone) e os potenciais motoristas são um exemplo de informação que pode ser incluída no cálculo de métricas de similaridade.
   - b) Na aprendizagem baseada em instâncias a generalização realizada pelo modelo de ML é baseada em métricas de similaridade, enquanto a aprendizagem baseada em modelos formula equações matemáticas que são usadas para fazer a generalização.
     - Na classificação de um potencial motorista para os serviços de corrida (aplicativos de transporte privado como Uber) é possível aplicar modelos baseados em instâncias para classificação de potenciais motoristas como também a construção de classificadores a partir de modelos matemáticos usados para tal tarefa (classificação).
   - c) A aprendizagem baseada em modelos realiza sua generalização por meio de equações matemáticas, tal como o algoritmo K-NN, enquanto a generalização por instâncias se baseia em métricas de desempenho.
   - d) Não existem diferenças entre os modos de aprendizagem de máquina baseados em instâncias e modelos.

10. **Exercício 10 (Fundamentos de ML)**

Em vários livros e artigos científicos, além dos diversos materiais que encontramos na internet, vemos o uso intercambiável ou equivalente entre as expressões algoritmos de aprendizagem supervisionados, técnicas e/ou modelos de machine learning supervisionadas. A Figura abaixo apresenta uma terminologia adequada e em sintonia com diversas literaturas de alto nível da área e nos permite esclarecer os conceitos e diferenças entre essas expressões para o caso supervisionado.

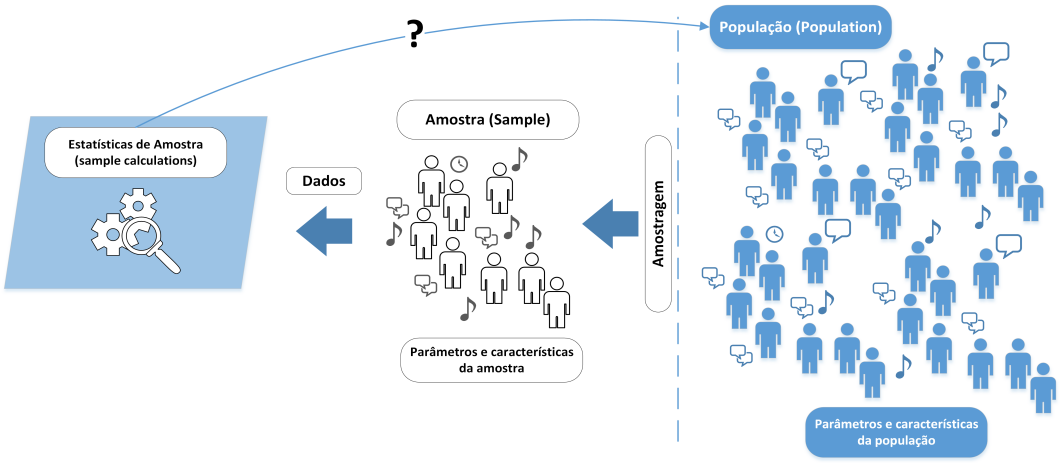


Baseado na figura, marque a alternativa que descreve o conceito de modelo supervisionado de machine learning:

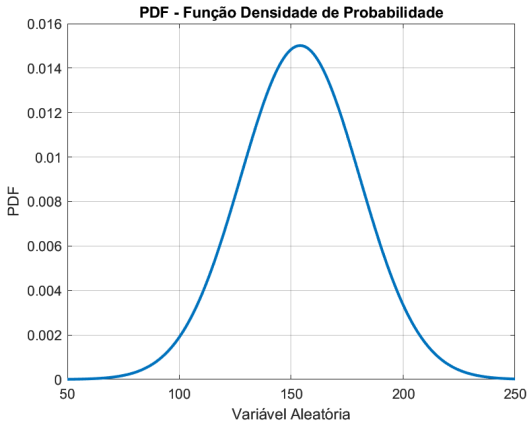
- a) O modelo de aprendizagem de máquina é o algoritmo de aprendizagem usado para treinamento.
- b) O modelo de aprendizagem de máquina supervisionado consiste na combinação entre uma função hipótese candidata e um algoritmo de aprendizagem.
- c) O modelo de aprendizagem de máquina consiste em um teste feito entre a função hipótese candidata e a verdadeira.
- d) O modelo de aprendizagem de máquina consiste no conjunto de funções hipóteses candidatas que são combinadas com um único algoritmo de aprendizagem de máquina.

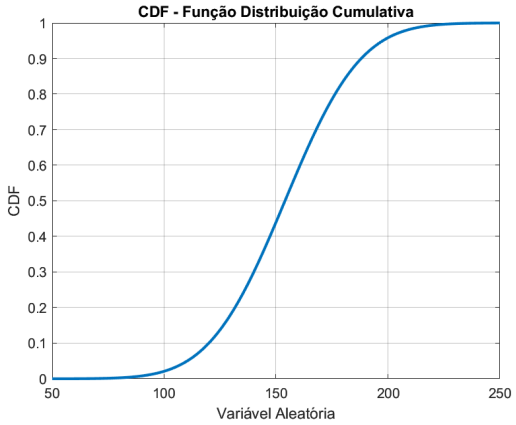
1.3 Exercícios de Revisão - Probabilidade/Estatística

1. Baseado na figura abaixo, apresente suas explicações sobre a diferença entre população e amostra. De que parte da estatística nós estamos falando quando queremos estimar parâmetros populacionais a partir dos dados de uma amostra?



2. Em probabilidade, nós utilizamos as funções PDF (probability density function) e CDF (cumulative distribution function), mostradas abaixo, para a caracterização estatística de variáveis aleatórias. Explique a diferença entre essas funções e como podemos calcular probabilidades a partir de cada uma delas (PDF e CDF).





1.3.1 Respostas - Exercícios de Revisão - Probabilidade/Estatística

1.4 Exercícios Computacionais

Os exercícios computacionais abaixo são voltados para os conhecimentos básicos de programação das diferentes linguagens que estamos estudando (e.g., R, Python, Matlab), incluindo conceitos introdutórios do aprendizado de máquina. Nosso objetivo é aumentar a familiaridade com os pacotes estatísticos e linguagens de programação estudados, aplicando-os para concretizar os conceitos introdutórios vistos em aula.

1. **Exercício Computacional - 1**
   Conceito explorado: Dataframes

   Dataframes podem ser interpretados como *tabelas de dados ou dados tabulados* e, talvez, sejam uma das estruturas de dados linha-coluna mais importantes na linguagem R no momento em que pensamos em aplicá-la para ciência de dados. Por isso, esse exercício tem um propósito básico: criar e explorar comandos básicos relacionados com dataframes no R.

	id	Empresa	Indices	Datas
1	1	A	500.3	2020-03-05
2	2	B	530.2	2020-04-21
3	3	C	630.5	2020-12-10
4	4	D	400.2	2020-10-15
5	5	E	940.2	2020-09-20

   - 1) Crie o dataframe mostrado na figura acima e armazene no objeto **df**
   - 2) Utilize a função **str**() e interprete os resultados sobre cada tipo de dado contido no dataframe
   - 3) Faça a extração apenas das colunas de empresas e índices
   - 4) Crie um array com os elementos relacionados com: a primeira (1) e terceira (3) linhas e a segunda (2) e quarta (4) colunas.
   - 5) Adicione uma nova coluna ao dataframe com os setores empresariais "IT", "adm", "executivo", "RH", "O&M" e armazene em novo dataframe chamado **df3**
   - 6) Combine o dataframe do item 1), dado por **df**, com o novo dataframe mostrado abaixo e armazene o resultado, também como dataframe, no objeto **dfn**. Estude as funções **rbind** e **cbind** para isso.

	id	Empresa	Indices	Datas
1	6	F	1200.3	2020-09-10
2	7	F	230.4	2020-07-08
3	8	G	100.5	2020-10-15
4	9	K	905.4	2020-06-07
5	10	L	1100.5	2020-02-22

2. **Exercício Computacional - 2**
   Conceito explorado: Geração de Números Aleatórios

   Gere uma amostra com 1000 observações que segue a distribuição de probabilidade Gaussiana com média $\mu = 10$ e desvio padrão $\sigma = 5$, Armazene os números aleatórios gerados no objeto **r**.

   - 1) Qual é o tipo de objeto **r**? Quais instruções você utilizou para verificar essa informação?
   - 2) Obtenha o histograma relacionado com o vetor **r**.
   - 3) Plote, sobre o histograma, a curva de densidade normal informando os valores de média e desvio padrão. Dica: no R, as funções **curve** e **dnorm** são úteis para solucionar esse ponto.
   - 4) Utilize o pacote **ggplot2** da linguagem R para obter o mesmo resultado dos itens anteriores.

3. **Exercício Computacional - 3**
Conceito explorado: Estatística Descritiva e Análise Exploratória de Dados

A Figura abaixo mostra um instrumento de teste (Field Fox Keysight) que pode ser usado em laboratório ou em campo para medições de sinais de radiofrequência como os presentes em sistemas de comunicações sem fio. Isso significa que podemos usar esse equipamento para análise de redes sem fio, cobertura de operadoras de telecomunicações, além de testes com dispositivos de RF e outros equipamentos de telecomunicações.

Nesse contexto, utilizamos esse equipamento para a realização de medições de intensidade de sinal no campus do Inatel a fim de levantar a cobertura de uma rede sem fio experimental, configurada para transmitir sinais na faixa de frequência de ondas milimétricas. O estudo de cobertura e propagação nessa faixa de frequência é um aspecto de pesquisa relevante para sistemas de comunicações da quinta geração de redes móveis. Nesse exercício, temos o objetivo de fazer a análise exploratória de dois conjuntos de dados, **dataset_1** e **dataset_2** exportados pelo instrumento de teste.

- 1) Faça a importação do arquivos **dataset_1** e **dataset_2** exportados pelo equipamento para o ambiente do RStudio.
- 2) Análise o resultado da importação, como as estruturas e tipos de variáveis. Quais são as principais informações contidas no arquivo?
- 3) Obtenha o histograma dos valores de potência de recepção coletados pelo equipamento em cada conjunto de dados.
- 4) Em qual localidade específica foram realizadas as medições de cada conjunto de dados?

4. **Exercício Computacional - 4**
Conceito explorado: Estatística Descritiva e Análise Exploratória de Dados

Considere o mesmo contexto do exercício anterior e um conjunto maior de arquivos .csv, que são exportados pelo instrumento de medida e armazenados em um diretório.

- 1) Com os arquivos .csv armazenados no diretório, elabore uma rotina em linguagem R para fazer a leitura de todos os arquivos de forma otimizada.
- 2) Capture os dados de geolocalização (latitude, longitude) de todos os arquivos, faça os processamentos e transformações necessárias, visando o armazenamento em um dataframe.
- 3) Apresente no mapa os dados de geolocalização obtidos no item anterior.

5. **Exercício Computacional - 5**
Conceito explorado: MSE - Mean Square Error

Considere o seguinte modelo de geração de dados mostrado abaixo:

$$\mathbf{y} = h(\mathbf{x}) + \varepsilon \tag{1.3}$$

Nesse modelo, $h(\mathbf{x}) = 3\mathbf{x} + 30$ consiste na função hipótese verdadeira, muitas vezes desconhecida na prática de ML, e ε é um termo que expressa a incerteza entre os valores da função hipótese verdadeira e a variável de saída ou resposta $\mathbf{y}$. Estatisticamente, ε é interpretado como um ruído, que nesse exercício segue a distribuição de probabilidade Gaussiana com média $\mu = 0$ e desvio padrão $\sigma = 15$. A notação em negrito usada ocorre em função de (1.3) ser um modelo vetorial de dados.

A variável explanatória $\mathbf{x}$ usada será um vetor de valores inteiros de zero (1) a cem (100). Logo, o modelo de dados é formado pelos vetores $\mathbf{y}$, $h(\mathbf{x})$ e ε, sendo cada um com dimensões (número de linhas e colunas) de 100×1.
Considere que um grupo de cientistas de dados já realizaram o trabalho de modelagem e encontraram uma função hipótese candidata dada por:

$$\hat{h}(\mathbf{x}) = 2.8\mathbf{x} + 32 \tag{1.4}$$

- 1) Construa esse modelo de geração de dados. Para que seja possível a reprodução de resultados em função do vetor aleatório ε utilize a semente (seed) 123 em seu código.
- 2) Faça um gráfico de dispersão da variável explanatória $\mathbf{x}$ com saída conhecida $\mathbf{y}$.
- 3) Obtenha o histograma relacionado com a variável de saída $\mathbf{y}$.
- 4) A equação do MSE, mostrada abaixo, é uma métrica de desempenho relacionada com qual tipo de tarefa de aprendizagem de máquina?

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{h}(\mathbf{x}_i))^2 \tag{1.5}$$

- 5) Faça a estimação do erro quadrático médio do modelo proposto pelos cientistas.
- 6) Faça uma análise: o modelo proposto é plausível para explicar os dados? De quais fatores esse desempenho depende?

6. **Exercício Computacional - 6**
Conceito explorado: MSE - Mean Square Error

Considere o mesmo modelo de geração de dados do exercício anterior. O objetivo aqui é constatar o impacto do desvio padrão σ sobre a performance do modelo proposto pelos cientistas de dados. Para isso, utilize a instrução abaixo, em linguagem R, para a geração de um vetor com diversos valores de desvio padrão para a incerteza Gaussiana retratada pelo termo ε.

```r
# Vetor com valores de desvio-padrão entre 0 e 20
std_vector = seq(1, 20, length.out = 100)
```

Para que seja possível explorar e visualizar o impacto de σ - realize, pelo menos, 1000 iterações do algoritmo. Especificamente, para cada valor de desvio padrão avaliado, armazene e faça o cálculo da média aritmética sobre 1000 valores de performance expressos pelo MSE. Um dica é utilizar estruturas em loop (for) para a implementação das iterações.

- 1) Construa o modelo de geração de dados incluindo as iterações para cada valor de σ
- 2) Faça um gráfico que mostra o impacto de σ, colocado sobre o eixo x, sobre o desempenho indicado pelo MSE, apresentado no eixo y.
- 3) Faça uma análise: o impacto com o aumento ou redução de σ é significativo para o modelo? Qual a justificativa?

7. **Exercício Computacional - 7**
Conceito explorado: MSE - Mean Square Error

Considere o mesmo modelo de geração de dados do exercício anterior. Agora, nosso objetivo é constatar o impacto do número de amostras n sobre a performance do modelo proposto pelos cientistas de dados. De forma similar ao caso anterior, utilize a instrução abaixo, em linguagem R, para a geração de um vetor com diversos valores de desvio padrão para a incerteza Gaussiana retratada pelo termo ε.

```
# Vetor com valores de número de amostras
n_vector = seq(10,100,5)
```

Para que seja possível explorar e visualizar o impacto de n - realize, pelo menos, 1000 iterações do algoritmo. Especificamente, para cada valor do número de amostras avaliado, armazene e faça o cálculo da média aritmética sobre 1000 valores de performance, expressos pelo MSE. Um dica é aproveitar as estruturas em loop (for) do exercício anterior para a implementação das iterações.

- 1) Construa o modelo de geração de dados incluindo as iterações para cada valor de n
- 2) Faça um gráfico que mostra o impacto de n, colocado sobre o eixo x, sobre o desempenho indicado pelo MSE, apresentado no eixo y.
- 3) Faça uma análise: o impacto com o aumento ou redução de n é significativo para o modelo? Qual a justificativa?

8. **Exercício Computacional - 8**
Conceito explorado: Introdução à Análise de Séries Temporais

Esse exercício tem o objetivo de explorar o assunto de séries temporais de forma introdutória. Para isso, existem alguns pacotes que podem nos auxiliar no objetivo desse exercício: capturar séries temporais do mercado financeiro e realizar sua visualização. Abaixo, estão listados três pacotes relacionados ao mercado financeiro que podem ser instalados e carregados na linguagem R.

- install.packages("quantmod")
- install.packages("xts")
- install.packages("moments")

Esses pacotes foram desenvolvidos exclusivamente para modelagem financeira quantitativa na linguagem R e permitem capturar séries temporais sobre as cotações de ações do mercado financeiro. Especificamente, estude e utilize a função **getSymbols** do pacote "quantmod" para obter séries temporais de diversas empresas presentes na bolsa de valores a partir de uma janela de tempo fornecida. Essa função consegue obter os dados diretamente das fontes "Yahoo Finance" (ainda ativo) e "Google Finance", que disponibilizam os dados gratuitamente. Utilize o nome "yahoo" para designar a fonte de dados na função **getSymbols**.

- 1) Use a função **getSymbols** do pacote quantmod para capturar as cotações de ações da empresa Petrobras de janeiro/2020 até os dias atuais.
- 2) Use a função **candleChart** do pacote quantmod para fazer a visualização da série temporal das cotações fechadas da Petrobras no período considerado. Pesquise e explore essa função; esse resultado é chamado de gráfico de velas (amplamente conhecido e usado em análises do mercado financeiro).
- 3) Explore a série temporal obtida da empresa, teste outros períodos de tempo e identifique o significado dos campos trazidos da série.
- 4) Use a função **addBands**, também do pacote quantmod, para plotar diretamente limites superior/inferior sobre a série temporal do item 1). Podemos parametrizar a função fornecendo: i) o período da média móvel e ii) a quantidade de desvio padrão relacionados com os limites.
- 5) A partir os itens anteriores, faça uma análise dos preços de cotações da Petrobras e verifique os reflexos do momento atual que estamos passando (março/abril/2020).