Programa de Pós-graduação em Genética
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Laboratório Nacional de Computação Científica

Laboratório de Bioinformática

Introdução a Aprendizagem de máquina para Bioinformática

Aulas

Msc. Amanda Araújo Serrão de Andrade
Biomédica, Mestre em Modelagem Computacional e Doutoranda em Genética
Orientadora: Dra. Ana Tereza Ribeiro de Vasconcelos

Apresentação

- Amanda Araújo Serrão de Andrade

- Biomédica (Faculdade Integrada Brasil Amazônia)
- Mestre em Modelagem Computacional pelo Laboratório Nacional de Computação Científica
- Doutoranda em Genética pela Universidade Federal do Rio de Janeiro

- **“Classificação de Arbovírus e Vírus vetor-específico utilizando algoritmos de Aprendizagem de Máquina”**

Overview do minicurso

Aula 1
• Obtenção dos conjuntos de dados • Representações numéricas de sequências biológicas • Redução da dimensionalidade do conjuntos de dados • Seleção de características
Aula 2
• Desbalanceamento • Validação cruzada • Classificação • Clusterização e Regressão
Aula 3
• Principais métricas de avaliação • Resultados • Comparação de performance

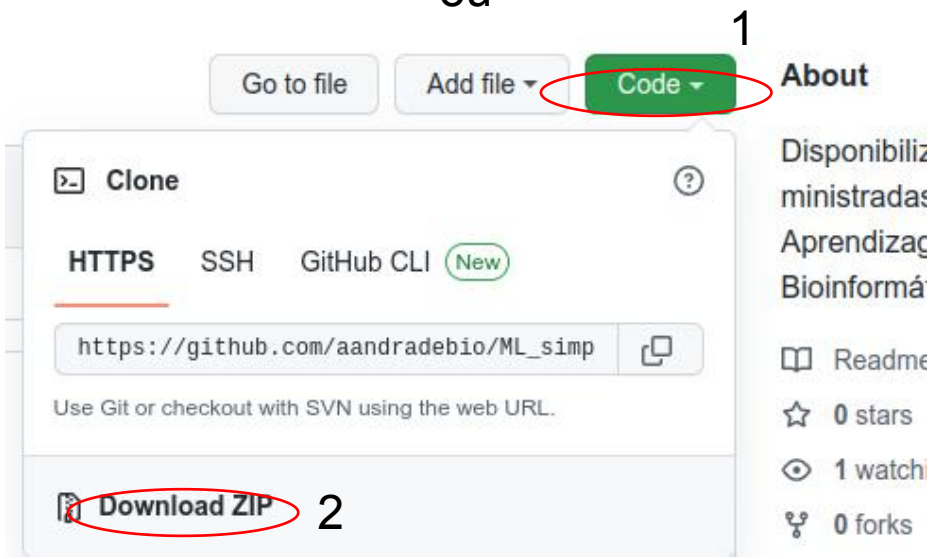
Dependências e repositório do curso

- Foi criado um repositório que contem todas as aulas práticas e conteúdos adicionais relacionado a este minicurso.

- https://github.com/aandradebio/ML_simposioPGGEN

$ wget https://github.com/aandradebio/ML_simposioPGGEN/archive/refs/heads/main.zip

ou



Dependências e repositório do curso

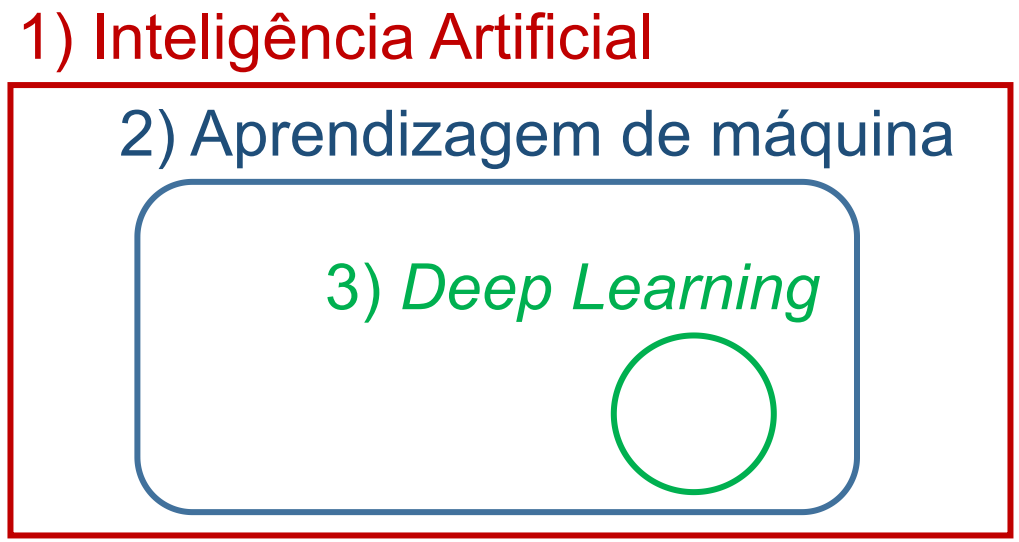
- Abram o RStudio na pasta descompactada do repositório.

- Dependências através do script install_dependencies.R depositado no GitHub ou através do comando abaixo:

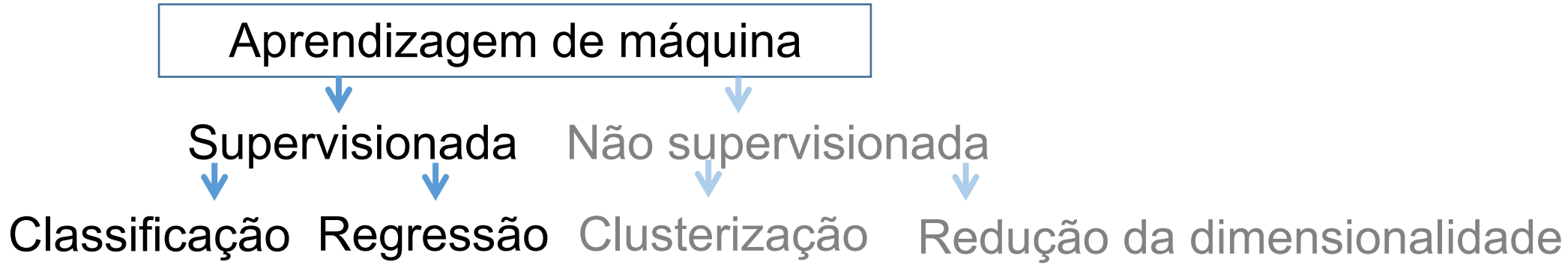
- install.packages(c("seqinrR","kmer", "caret", "MLeval", "ggplot2", "ggtree", "dplyr", "ape", "tidyverse","e1071", "randomForest", "ranger", "tidyr", "adabag", "extraTrees","ISLR","caretEnsemble"))

Inteligência artificial

- Inteligência Artificial: a capacidade das máquinas de aprender e decidir, sem interferência humana.
- A Aprendizagem de máquina é um dos campos da Inteligência Artificial, utiliza algoritmos com a finalidade de extrair informações de dados brutos, representá-los por meio de algum tipo de modelo matemático e fazer predições.
- A *Deep Learning* utiliza algoritmos inspirados na arquitetura natural do cérebro humano aprendendo com grandes quantidades de dados.



Os principais tipos de Aprendizagem de máquina

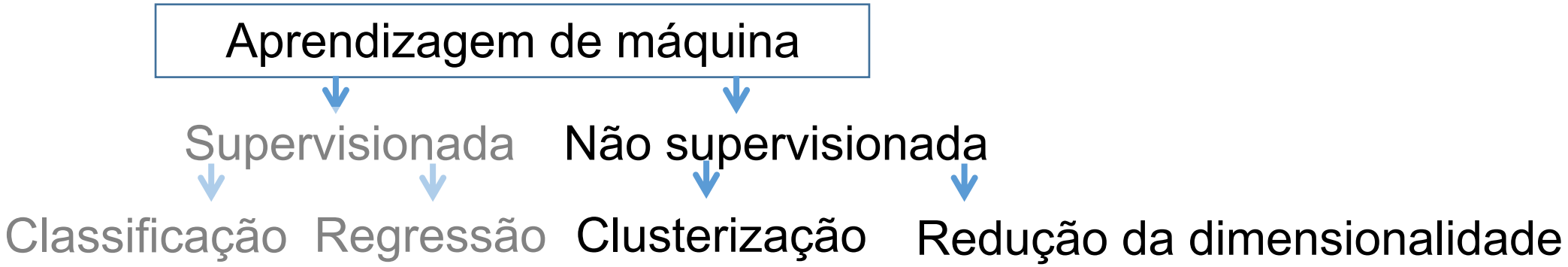


- Supervisionada: o modelo é construído a partir dos dados de entrada, apresentados na forma de pares ordenados (instância — classe desejada). Dizemos que estes dados são rotulados, pois sabemos de antemão a classe esperada para cada instância de dados.

- Classificação: Classes categóricas. Essa classificação pode ser binária (duas classes, 1 ou 0) ou multiclasse (três ou mais classes). Exemplo: prever se um tumor é maligno ou benigno.

- Regressão: o modelo deve prever uma classe a partir de uma faixa contínua de valores possíveis. A exatidão de um algoritmo de regressão é calculada com base na variação entre a classe precisa e a classe prevista. Exemplo: prever a idade

Os principais tipos de Aprendizagem de máquina



- Não supervisionada: consiste em treinar uma máquina a partir de dados que não estão rotulados e/ou classificados. Os algoritmos encontram padrões complexos entre as instâncias. Utiliza medidas de similaridade e de diferenças.

- Clusterização: Agrupamento de instâncias similares, de tal forma que elementos em um cluster compartilhem um conjunto de propriedades comuns que os diferencie dos elementos de outros clusters.

- Redução da dimensionalidade: Identificar a correlação entre diferentes instâncias de um conjunto de dados. Visualização da estrutura dos dados.

Qual algoritmo utilizar de acordo com o problema de pesquisa?

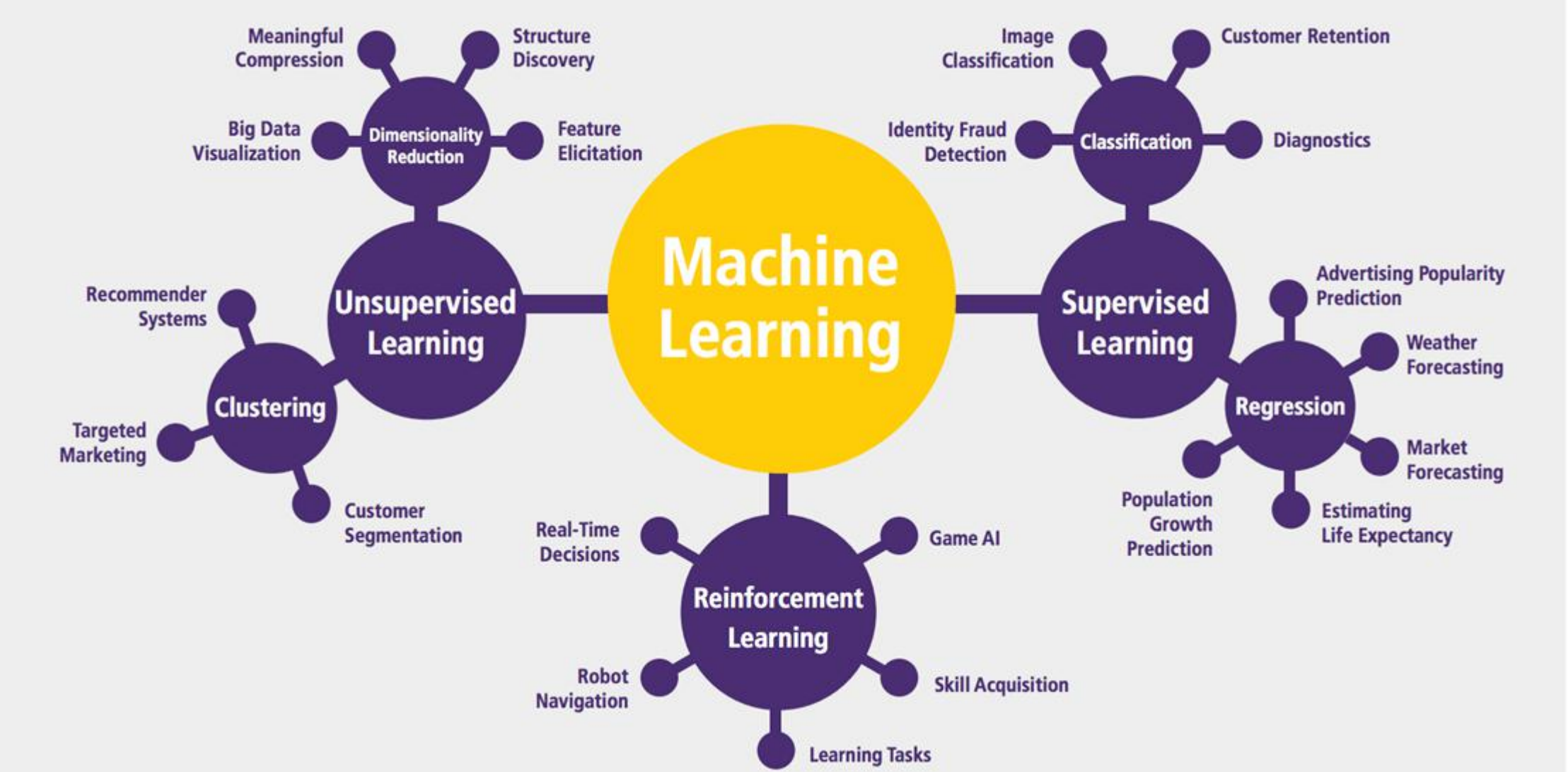


Fig 1. Algoritmos de Aprendizagem de Máquina e suas aplicações.Fonte: www.malicksarr.com/how-to-choose-the-right-algorithm-in-machine-learning/

Qual algoritmo utilizar de acordo com o problema de pesquisa?

Dever de casa: Dê exemplos de trabalhos na área da Bioinformática para os diferentes tipos de Aprendizado.

Classificação supervisionada: Identificar se um vírus desconhecido infecta plantas ou mosquitos?

Regressão supervisionada: Identificar a probabilidade da ocorrência de recombinação entre diferentes sequências (Regressão logística)

Clusterização: Alinhamento global de sequências para identificar proximidade entre elas (Clustering hierárquico)

Redução da dimensionalidade: Identificação de *motifs* capazes de discriminar a proteína L de diferentes vírus segmentados (Engenharia de características)

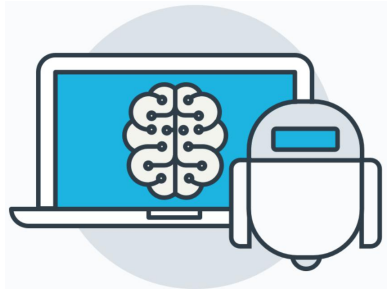
Exemplo prático: Como identificar se um virus desconhecido infecta plantas ou mosquitos?

1) Amostra biológica

2) Sequenciamento

3) Representação numérica

4) Pré-processamento e Redução da dimensionalidade



5) Algoritmo de Aprendizagem

Modelo preditivo

6) Predição de dados desconhecidos

Aprendizagem de máquina a partir de sequências biológicas

- Todos os dados de entrada devem ser numéricos, sejam provenientes de sequências biológicas ou não.

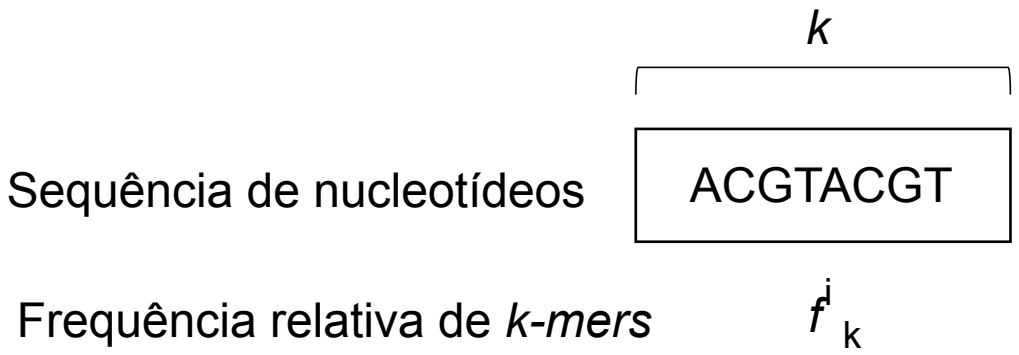
- Existem diferentes representações numéricas:

- *K-mers*;
- Codificação de janelas das sequências;
- Obtenção e codificação da estrutura secundária do RNA;
- One-hot-encoding (resultado binário);
- *Frequency Chaos Game Representation;*

- Vantagens e desvantagens.

Aprendizagem de máquina a partir de sequências biológicas

- Os *k-mers* são subsequências de nucleotídeos que apresentam tamanho *k*;
- O número de ocorrência dos *k-mers* é utilizado para a obtenção das suas frequências relativas;
- Quanto maior o k, maior o custo computacional;



3) Representação numérica

Matriz de características (*feature Matrix*)

- Dados de entrada para os algoritmos; Base para a aprendizagem;
- Os vetores resultantes são as representação numérica das características encontradas em uma sequência de nucleotídeos;

Exemplo de matriz de características

	ID	AAA	AAC	AAG	....	TTG	TTT	Classes
Instâncias	NC_58966	0.07	0.05	0.00	-	0.00	0.01	plant.vir
	NC_78544	0.06	0.02	0.00	-	0.00	0.11	mosquito.vir
	NC_96874	0.04	0.08	0.00	-	0.10	0.20	plant.vir

Características

Informações sobre as instâncias

Matriz de características (*feature Matrix*)

- Dados de entrada para os algoritmos; Base para a aprendizagem;
- Os vetores resultantes são as representação numérica das características encontradas em uma sequência de nucleotídeos;

Essencial para a aprendizagem supervisionada

Exemplo de matriz de características

	ID	AAA	AAC	AAG	....	TTG	TTT	Classes
Instâncias	NC_58966	0.07	0.05	0.00	-	0.00	0.01	plant.vir
	NC_78544	0.06	0.02	0.00	-	0.00	0.11	mosquito.vir
	NC_96874	0.04	0.08	0.00	-	0.10	0.20	plant.vir

Características

Informações sobre as instâncias

Continuação

Exemplo prático: Como identificar se um virus desconhecido infecta plantas ou mosquitos?

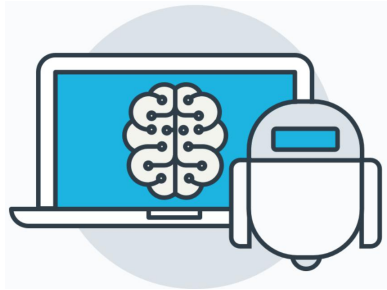
1) Amostra biológica

2) Sequenciamento

3) Representação numérica

4) Pré-processamento e Redução da dimensionalidade

5) Algoritmo de Aprendizagem



Modelo preditivo

6) Predição de dados desconhecidos

A maldição da dimensionalidade

- A Maldição da Dimensionalidade é um fenômeno que aparece quando temos uma grande quantidade de dados e muitas características para um mesmo problema.

- Relacionado a maiores custos computacionais;
- Menor poder preditivo do modelo treinado;
- *Overfitting* (o modelo erra a predição de dados desconhecidos);

- Então, para evitar isso, utilizamos técnicas de Redução de Dimensionalidade.

Pré-processamento e redução da dimensionalidade

- Por quê utilizar estas técnicas?

Simplificar os dados, remover ruídos, facilitar o processamento computacional e visualização.

Podem ser aplicadas as características numéricas e categoricas.

Filtro de acordo com a pergunta biológica (exemplo: remoção de certas famílias virais);

Remoção de dados duplicados (pode enviesar o modelo preditivo na etapa de teste com novos dados);

Remoção e substituição de dados indisponíveis (NA or NAN);

Remoção de características altamente correlacionadas (entender como as características se relacionam entre si e se existe redundância);

Pré-processamento e redução da dimensionalidade

- Seleção de características com as funções varIMP e RFE do pacote Caret. Medida em proporções, quão importante aquela característica é para o modelo, onde quanto maior o valor, mais importante (Prática)

- A importância de uma característica é calculada a partir do impacto causado na acurácia da classificação pela sua remoção.

- O pré-processamento da matriz de característica depende do algoritmo que pretende usar. Algumas características podem ser importantes para a classificação e dispensáveis para a regressão, por exemplo.

Desbalanceamento de dados

- Desproporção entre as instâncias de diferentes classes.
- O desbalanceamento influencia negativamente a classificação, pois o algoritmo tende a classificar a classe majoritária.
- Exemplo: comparação entre classes de vírus mais estudados e menos estudados.

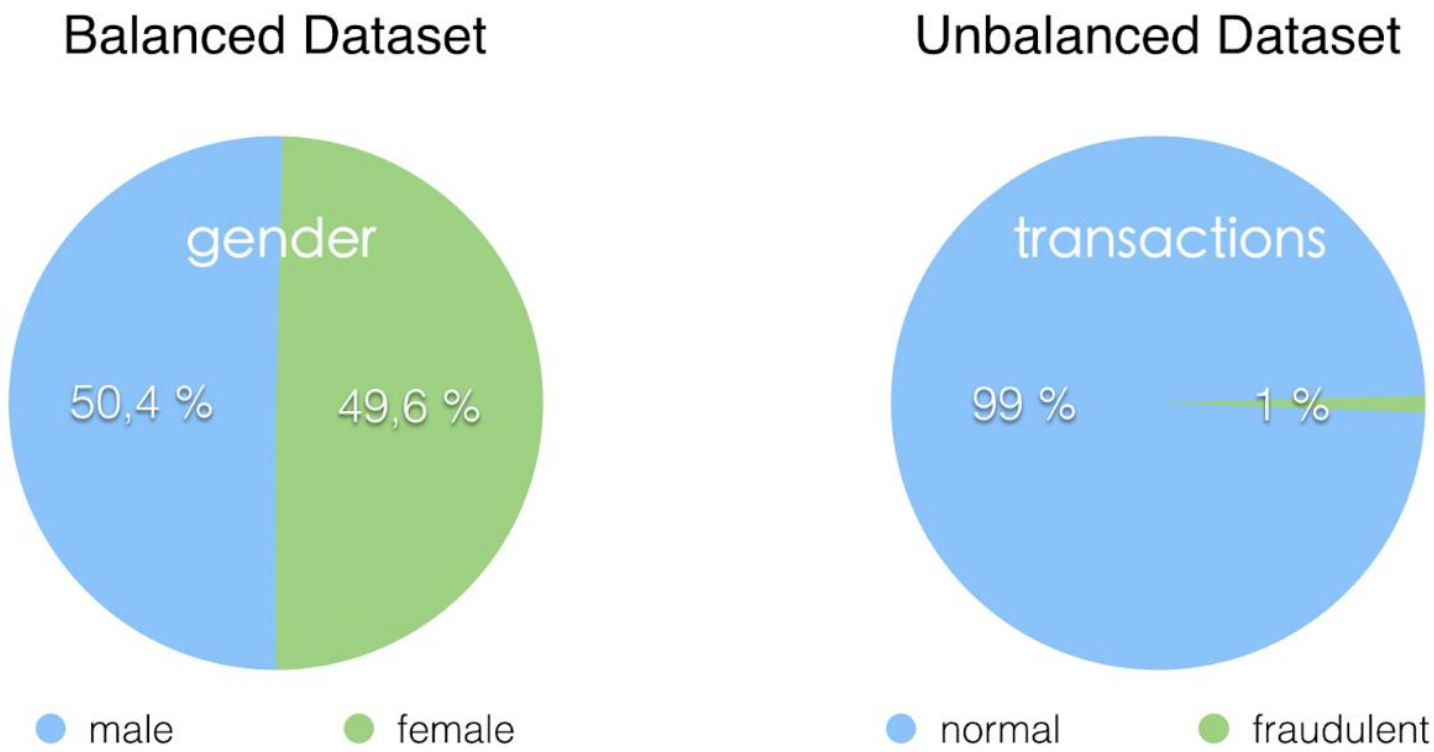


Fig 2. Exemplo de conjunto de dados desbalanceados. Fonte: https://blog.strands.com/unbalanced-datasets

Como lidar com uma matriz de características desbalanceada?

- *Undersampling* (Prática)*;*
- *Oversampling* (Prática)*;*
- *Random resampling* (Prática);
- Gerar dados artificialmente;

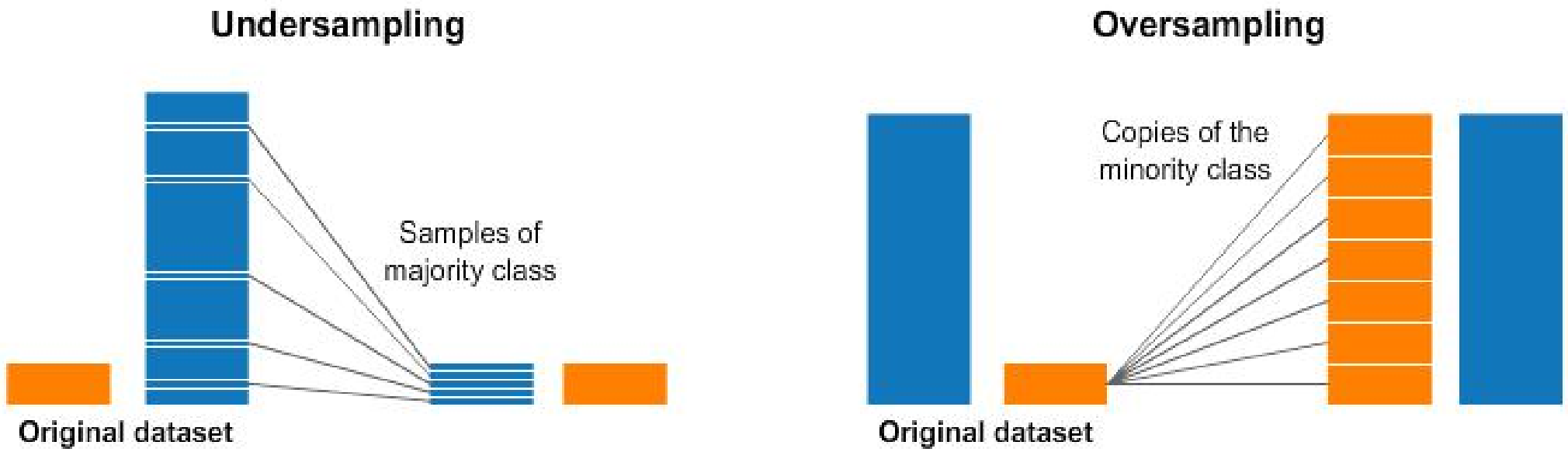


Fig 2. Principais estratégias para lidar com dados desbalanceados. Fonte:https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Referências e material de apoio

- https://machinelearningmastery.com/machine-learning-in-r-step-by-step/
- https://topepo.github.io/caret/
- https://github.com/aandradebio/ML_simposioPGGEN
- https://www.dataquest.io/blog/tutorial-getting-started-with-r-and-rstudio/
- https://rpubs.com/DeclanStockdale/799284

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

- machinelearningmastery.com

Dúvidas?

aandradebio@gmail.com

- As dúvidas mais específicas devem vir acompanhadas com código e um conjunto de dados de teste para agilizar a resposta

Obrigada!!

aandradebio@gmail.com

atrv@lncc.br