Aprendizagem de Máquina

Conceitos Fundamentais: o Retorno

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA

Sumário

Definições

Tipos de Aprendizagem de Máquina
- Aprendizagem Supervisionada
- Aprendizagem não-Supervisionada
- Aprendizagem Semi-Supervisionada
- Aprendizagem por Reforço

Os Benditos Dados
- Conjuntos de Dados
- Pré-processamento

O que é Aprender?

- Ganhar conhecimento através do estudo, experiência ou sendo ensinado
- **Aprendizagem** X **Aprendizado**
  - **Aprendizagem** é o processo pelo qual se adquire o conhecimento $\rightarrow$ Algoritmos
  - **Aprendizado** é o conhecimento adquirido $\rightarrow$ Modelos
- Na disciplina de **Aprendizagem** de Máquina, focamos no estudo de **algoritmos** para adquirir descrições estruturais (**modelos**) sobre exemplos de dados

Aprendizagem de Máquina

- Os algoritmos da Aprendizagem de Máquina permitem que escrevamos programas cujo desempenho tende a melhorar à medida que **“ganham experiência”**
- Essa experiência corresponde aos dados que são fornecidos ao programa
- Os algoritmos buscam extrair **hipóteses** dos dados
  - A probabilidade de chover dado um valor de umidade relativa do ar x é p

Inferência Indutiva

- Processo para conclusão sobre o todo por meio do exame de apenas alguns membros
- Raciocínio do particular pra o geral
- Exemplo:
  - Todos os pacientes com TDAH em 1986 sofriam de ansiedade
  - Todos os pacientes com TDAH em 1987 sofriam de ansiedade

  ⋮

  - Posso inferir que “Todos os pacientes que sofrem de TDAH também sofrem de ansiedade”
- Cada algoritmo de Aprendizagem de Máquina possui um viés indutivo diferente

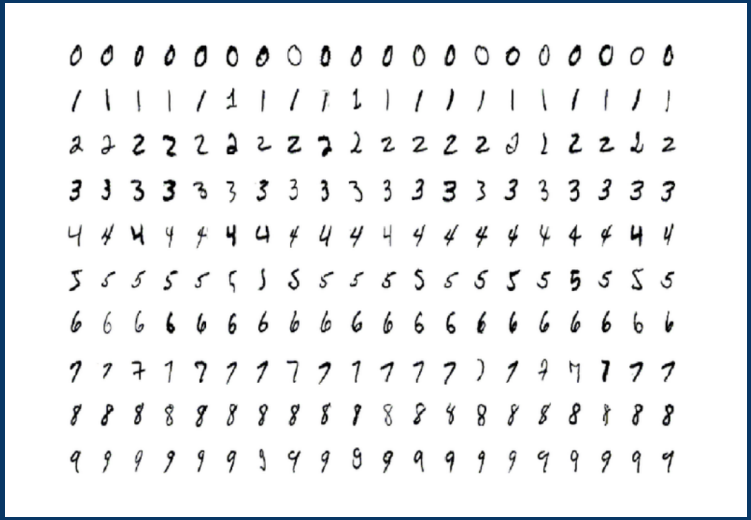
Exemplos de Aplicações

- A partir de informações relativas a gravidez, aprender a prever classes de futuras pacientes de alto risco que devem fazer cesárea
- Prever a chance de um cliente deixar de consumir certo produto (*churn prediction*)
- Recomendar filmes

Tarefa, Medida e Experiência

- De uma maneira geral, o uso da Aprendizagem de Máquina para solucionar uma tarefa envolve:
  - Otimizar a realização de uma tarefa T
    - Reconhecer e classificar caracteres manuscritos
  - Em relação a uma medida de desempenho P
    - Porcentagem de caracteres classificados corretamente
  - Baseada na experiência E
    - Base de dados de caracteres manuscritos com a respectiva classificação

Exemplo



Tipos de Aprendizagem de Máquina

Tipos de Aprendizagem de Máquina

- Aprendizagem supervisionada
- Aprendizagem não-supervisionada
- Aprendizagem semi-supervisionada
- Aprendizagem por reforço

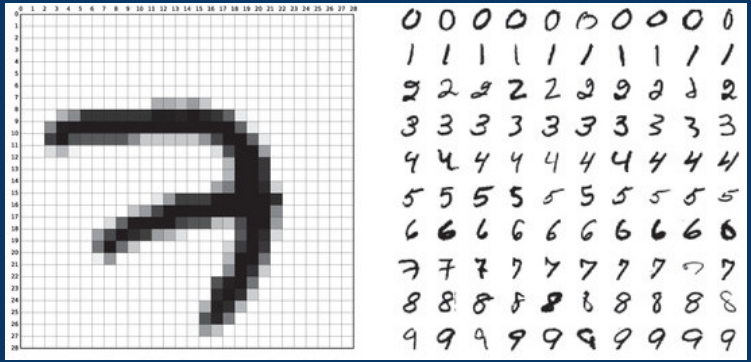
Aprendizagem Supervisionada

- O algoritmo de aprendizagem recebe um conjunto de exemplos/instâncias de ajuste/treinamento em que um rótulo alvo é conhecido
- As tarefas incluem:
  - Classificação: determinar a classe de uma instância dados os seus valores descritivos/atributos, i.e. $\hat{y} = \arg\max_y P(Y = y|X = \mathbf{x})$
  - Regressão: estimar o valor esperado da variável alvo de uma instância dados os seus atributos, i.e. $\hat{y} = \mathbb{E}[Y|X = \mathbf{x}]$
- Cada exemplo é descrito por um vetor de valores (atributos) e pelo rótulo (classe ou valor alvo) associado

Aprendizagem Supervisionada

- Os algoritmos de classificação podem ser divididos ainda em:
  - **Generativos:** dadas as variáveis X e Y, o objetivo é encontrar a distribuição de probabilidade conjunta $P(X, Y)$ para a partir daí determinar $P(Y|X = \mathbf{x})$
    - Naïve Bayes
    - Discriminante linear
  - **Discriminativos:** buscam estimar diretamente a probabilidade condicional $P(Y|X = \mathbf{x})$ (regressão logística) ou nem assumem modelos probabilísticos (perceptron, SVM)

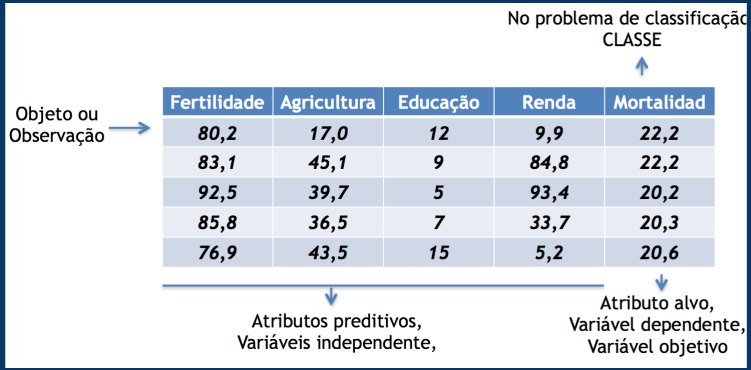
Exemplo 1



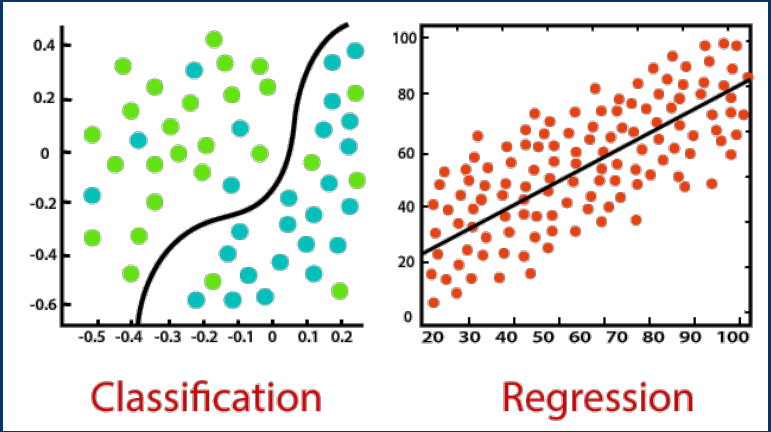
Exemplo 2

Tamanho (P)	Largura (P)	Tamanho (S)	Largura (S)	Espécie
5,1	*3,5*	*1,4*	*0,2*	*Setosa*
4,9	*3,0*	*1,4*	*0,2*	*Setosa*
7,0	*3,2*	*4,7*	*1,4*	*Versicolor*
6,4	*3,2*	*4,5*	*1,5*	*Versicolor*
6,3	*3,3*	*6,0*	*2,5*	*Virginica*
5,8	*2,7*	*5,1*	*1,9*	*Virginica*

Exemplo 3


No problema de classificação
CLASSE
Objeto ou
Observação
Fertilidade
Agricultura
Educação
Renda
Mortalidad
80,2
17,0
12
9,9
22,2
83,1
45,1
9
84,8
22,2
92,5
39,7
5
93,4
20,2
85,8
36,5
7
33,7
20,3
76,9
43,5
15
5,2
20,6
Atributos preditivos,
Variáveis independente,
Atributo alvo,
Variável dependente,
Variável objetivo

Exemplo 4

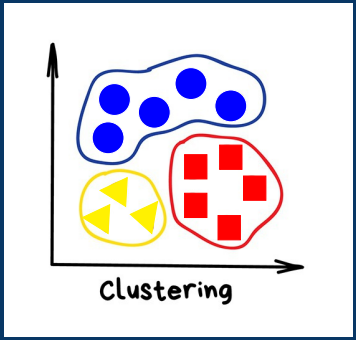

0.4
0.2
0
-0.2
-0.4
-0.6
-0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2
Classification
100
80
60
40
20
0
20 30 40 50 60 70 80 90 100
Regression

Aprendizagem não-Supervisionada

- O indutor analisa os exemplos fornecidos e tenta determinar se existem padrões previamente desconhecidos nos dados
  - Comumente, o objetivo é encontrar $P(X)$, ou seja, a distribuição de probabilidade de X
- Aqui, os dados não possuem rótulos para ajudar no ajuste dos modelos
- Tarefas não-supervisionadas incluem:
  - Agrupamento
  - Detecção de anomalias
  - Modelos de variáveis latentes

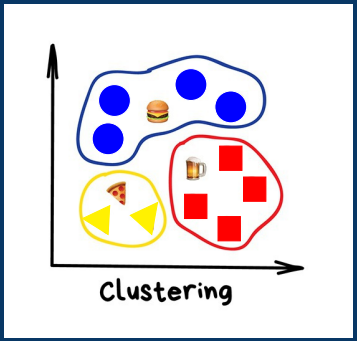
Agrupamento

- O algoritmo verifica se as instâncias observadas podem ser arranjadas de alguma maneira, formando grupos (*clusters*)
- O objetivo é que os clusters sejam maximamente parecidos internamente e maximamente diferentes entre si
- Após a determinação dos grupos, é necessário analisá-los para entender o que cada um representa



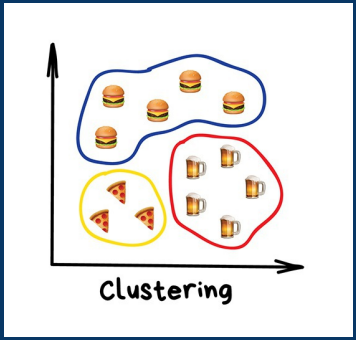
Agrupamento

- O algoritmo verifica se as instâncias observadas podem ser arranjadas de alguma maneira, formando grupos (*clusters*)
- O objetivo é que os clusters sejam maximamente parecidos internamente e maximamente diferentes entre si
- Após a determinação dos grupos, é necessário analisá-los para entender o que cada um representa



Agrupamento

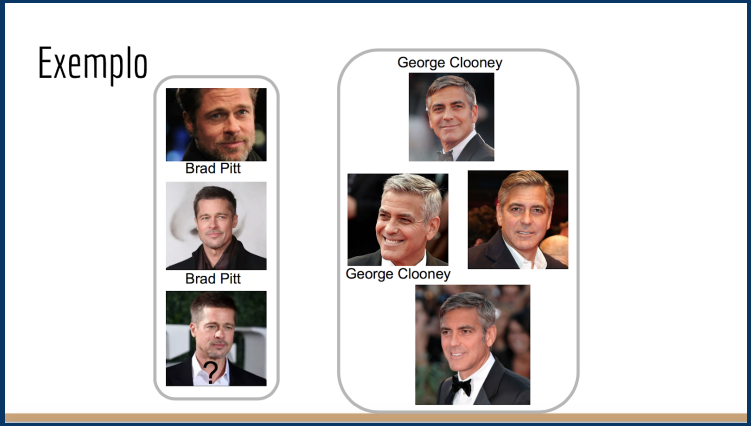
- O algoritmo verifica se as instâncias observadas podem ser arranjadas de alguma maneira, formando grupos (*clusters*)
- O objetivo é que os clusters sejam maximamente parecidos internamente e maximamente diferentes entre si
- Após a determinação dos grupos, é necessário analisá-los para entender o que cada um representa



Aprendizagem Semi-Supervisionada

- Esse tipo de aprendizagem assume que o conjunto de treinamento possui instâncias rotuladas e (frequentemente muito mais) instâncias não-rotuladas
- O objetivo dos algoritmos semi-supervisionados é usar toda a informação possível,
- Em notação, usa-se as instâncias rotuladas para estimar $P(Y|X)$ e todas as instâncias, incluindo as não rotuladas, para estimar $P(X)$, tudo isso simultaneamente, de forma que uma estimativa ajude a outra

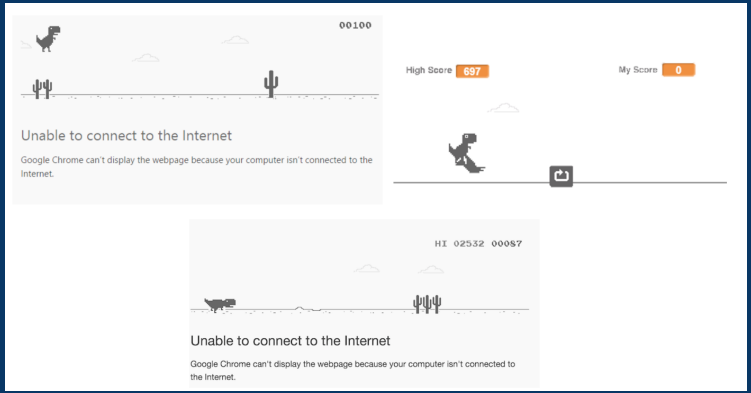
Exemplo



Aprendizagem por Reforço

- Aprendizagem por reforço (Sutton, R.S. e Barto, A.G., 1998) envolvem situações em que um ou mais agentes aprendem por tentativa e erro ao atuar sobre um ambiente dinâmico
- Não há uma fonte externa de exemplos. Há apenas a própria experiência do agente
- É necessário definir que ações o agente pode desempenhar e qual é a medida de desempenho

Exemplo


00100
Unable to connect to the Internet
Google Chrome can't display the webpage because your computer isn't connected to the Internet.
High Score 697
My Score 0
HI 02532 00087
Unable to connect to the Internet
Google Chrome can't display the webpage because your computer isn't connected to the Internet.

Os Benditos Dados

Explicando os dados

- Atributos podem ser físicos ou abstratos, como sintomas
- Cada objeto/instância é descrito por um conjunto de atributos de entrada ou vetor de características
- Cada objeto corresponde a uma ocorrência/observação
- Os atributos estão associados a propriedades dos objetos

Conjuntos de dados

- Os dados que usamos para treinar nossos modelos são agregados em um **conjunto ou base de dados** (*data set* ou *dataset*)
- O conjunto de dados costuma ser representado por uma matriz $\mathbf{X}_{n \times d}$
  - n é o número de instâncias
  - d é o número de atributos de cada instância e define a dimensionalidade do espaço do problema

Exemplo

```
df = pd.read_csv('hospitall.csv', delimiter=";")

df
```

	Identificador	Nome	Idade	Sexo	Peso	Manchas	Temperatura	Internacoes	Estado	Diagnostico
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	38.0	3	RJ	Doente
6	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
7	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

Conjuntos hospital

- Este conjunto aparentemente tem $d = 10$ variáveis
- No entanto, a primeira e a segunda são apenas identificadores de paciente
- E a última, que indica o diagnóstico, possivelmente sera selecionada como alvo em uma tarefa de classificação, sendo tratada como uma variável separada Y
- Portanto, $d = 7$

Pré-processamento

- Antes de alimentar um algoritmo de aprendizagem de máquina com o conjunto de dados observados, comumente precisamos realizar diversas atividades de preparação dos dados, incluindo:
  - Eliminação manual de atributos
  - Integração de dados
  - Amostragem
  - Balanceamento
  - Limpeza
  - Redução de dimensionalidade
  - Transformação

Pré-processamento

- Essas técnicas são usadas para melhorar a qualidade dos dados, i.e. tornar mais fácil o ajuste de modelos
- Minimizam problemas de ruídos, anomalias/outliers, valores/rótulos incorretos, duplicados ou ausentes
- Também podem adequar os dados para uso de determinados algoritmos, e.g. algoritmos com entradas exclusivamente numéricas

Eliminação manual de atributos

- Removemos atributos que não contribuem para a construção dos modelos
- Nesse momento, o conhecimento e a experiência dos especialistas são fundamentais

Removendo colunas

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv("hospital.csv", sep = ';')

df = df.drop(columns=['identificador','nome'])

df
```

	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	28	M	79	Concentradas	38.0	2	SP	Doente
1	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	49	M	92	Espalhadas	38.0			
3	18	M	43	Inexistentes	38.5			
4	21	F	52	Uniformes	37.6			
5	22	F	72	Inexistentes	58.0			
6	19	F	87	Espalhadas	39.0			
7	34	M	67	Uniformes	38.4			

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
6	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
7	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

Integração de dados

- Essa atividade trata da junção de duas bases de dados que possuem informações sobre os mesmos objetos
- Devemos buscar atributos comuns nos conjuntos que serão combinados
  - Exemplos: CPF, CNPJ e identificadores de uma maneira geral, além de outros atributos que podem estar repetidos
- Atributos cruzados devem ter um valor único para cada objeto

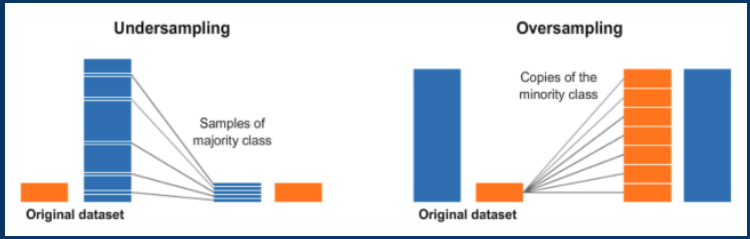
Amostragem de dados

- Alguns algoritmos de aprendizagem de máquina podem ter dificuldade de lidar com grandes volumes de dados
- Assim, torna-se útil obter uma amostra **representativa** dos dados para treinar o modelo
  - Os dados da amostra devem seguir a mesma distribuição dos dados originais (**qual?**)
- Diferentes amostras podem gerar modelos diferentes (mais sobre isso na aula sobre avaliação de modelos)

Balanceamento de dados

- Em certas aplicações (como na medicina), é comum que uma classe seja muito mais frequente do que outra
- Nesses casos, o modelo de AM pode aprender a “chutar” sempre a classe mais frequente
- Soluções:
  - Equalizar os tamanhos das classes
    - Subamostragem (*undersampling*)
    - Sobreamostragem (*oversampling*)
  - Classificação baseada em custos (mais sobre isso nos tópicos adicionais no final do curso)
  - Ajustar um modelo por classe

Balanceamento de dados




```
#Randommicamente seleciona 4 instâncias a partir da classe 'Doente'
df.loc[df['diagnostico'] == "Doente"].sample(n=4,random_state=2)
```

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente

https://imbalanced-learn.readthedocs.io/en/stable/over_sampling.html#smote-adasyn

Limpeza dos dados

- Remove problemas relacionados à qualidade dos dados
- Dados ruidosos: erros de registro, variações de qualidade de sinal
  - Diferente de outliers
- Inconsistentes: contradizem valores de outros atributos do mesmo objeto
- Redundantes: dois ou mais objetos/atributos com os mesmos valores
- Incompletos (com ausência de valores)

Limpeza dos dados

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	NaN	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67.0	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92.0	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43.0	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52.0	Uniformes	NaN	1	PE	Doente
5	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
6	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
8	1322	Marta	19	F	87.0	Espalhadas	39.0	6	AM	Saudavel
9	3027	Paulo	34	M	67.0	Uniformes	38.4	2	GO	Saudavel

Dados incompletos

- Possibilidades de correção:
  - Eliminar instâncias/colunas com valores ausentes
  - Usar média/moda/mediana dos valores conhecidos
  - Criar um novo valor que indique o atributo tem valor faltante
  - Estimar a distribuição conjunta dos atributos para depois preencher os faltantes com os valores mais prováveis
  - Usar algoritmos capazes de lidar com dados ausentes

Dados inconsistentes

- Problemas na anotação dos dados podem resultar em atributos de entrada que não explicam o atributo alvo/classe

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	67	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Doente
5	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Saudavel
6	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
8	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
9	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
10	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

Dados redundantes

- O mesmo atributo pode aparecer em dois formatos diferentes: idade X data de nascimento (string ou colunas numéricas)
- Atributos podem ser altamente correlacionados
  - Não há acréscimo de informação ao manter os dois
  - Mantém-se apenas um
  - Boa parte dos algoritmos de AM assume que não há correlação entre atributos

Outliers

- Dados que diferem bastante dos outros elementos do conjunto de dados ou de sua classe
- Podem ser retirados ou mantidos, caso deseje-se gerar modelos que modelam a sua existência
- Existem técnicas cujo objetivo é detectar outliers (veremos mais à frente no curso)

Transformação de dados

- Frequentemente é necessário transformar os tipos ou valores dos atributos para obter um melhor ajuste dos modelos
- Pode-se discretizar valores numéricos ou transformá-los em intervalos
- Pode-se transformar atributos categóricos com p categorias em p atributos binários
  - One-hot encoding, variáveis dummy
- E fazemos também a conhecida normalização, quando os atributos têm escalas muito diferentes

$$X_{novo} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Novos valores entre $[0, 1]$

$$Z = \frac{X - \mu}{\sigma}$$

Lida melhor com outliers

Aprendizagem de Máquina

Conceitos Fundamentais: o Retorno

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA