Introdução a machine learning com aplicações em R

Aula 2 - Pré-processamento dos dados

Felipe Barletta

Departamento de Estatística

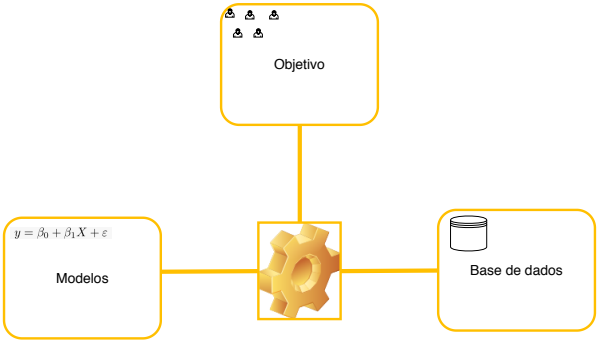
03 novembro, 2020

Sumário

Revisão - *Machine Learning*

- Intersecção de métodos estatísticos, Ciência da Computação (por meio de dados)
- É a tarefa de apredender com os dados e fazer generalizações

Revisão - Os três pilares



Revisão - Termos comuns

Machine Learning	Estatística	O que é?
Label	Variável resposta	Variável a ser predita - (Y)
Feature ou característica	Variável preditora	Variável que pode predizer (Y)
output	Resposta	Predição do modelo
input	Variável preditora	Variável que está no modelo
Variável binária	Variável dicotômica	Variável com duas classes
Treinamento	Ajuste do modelo	Estrutura matemática
Missing values	*Missing values*	Valores faltantes
Cross validation	*Cross validation*	Validação cruzada

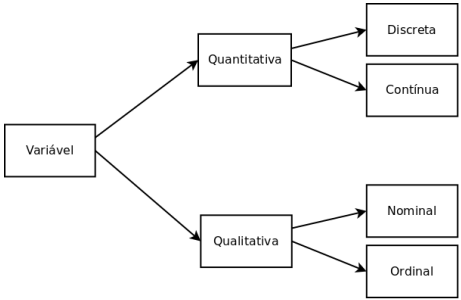
Revisão - Estrutura dos dados

Features ou características

duration.in.month	credit.amount	present.residence.since	foreign.worker	creditability	ID
6	1169	4	yes	good	1
48	5951	2	yes	bad	2
12	2096	3	yes	good	3
42	7882	4	yes	good	4
24	4870	4	yes	bad	5
36	9055	4	yes	good	6
24	2835	4	yes	good	7
36	6948	2	yes	good	8
12	3059	4	yes	good	9
30	5234	2	yes	bad	10

Exemplos

Tipos de características ou variáveis



- **Nominal:** São categorias que não podem ser ordenadas
- **Ordinal:** São categorias que seguem uma ordem natural
- **Discreta:** Seu suporte é um conjunto finito ou infinito enumerável
- **Continua:** Seu suporte é um conjunto infinito não-enumerável

Pré-processamento

- Formatar os dados.
- Retirar amostras
- Remover ou corrigir dados faltantes
- Agregar ou criar novas características(variáveis)
- Verificar associação entre características(variáveis)

Pré-processamento

- Formatar os dados.
- Retirar amostras
- Remover ou corrigir dados faltantes
- Agregar ou criar novas características(variáveis)
- Verificar associação entre características(variáveis)

Pré-processamento

- Formatar os dados.
- Retirar amostras
- Remover ou corrigir dados faltantes
- Agregar ou criar novas características(variáveis)
- Verificar associação entre características(variáveis)

Pré-processamento

- Formatar os dados.
- Retirar amostras
- Remover ou corrigir dados faltantes
- Agregar ou criar novas características(variáveis)
- Verificar associação entre características(variáveis)

Pré-processamento

- Formatar os dados.
- Retirar amostras
- Remover ou corrigir dados faltantes
- Agregar ou criar novas características(variáveis)
- Verificar associação entre características(variáveis)

Feature Engineering

É a arte de extrair informações dos dados coletados

1. Criação de características(variáveis)
2. Dados faltantes
3. Dados desbalanceados
4. Características(variáveis) correlacionadas

Exemplo: Desfecho de pacientes internados em um hospital

Desfecho	Quantidade
Ainda Internado	246800
Alta Médica	4000
Alta Melhorado	502
Alta	70
Atendendimento Cancelado	203
Evadiu-se	5
Internacão Cancelada	6
Internamento Cancelado	14
Óbito	400
Óbito - Pós-Operatório	29

Exemplo: Desfecho de pacientes internados em um hospital

Desfecho	Quantidade
Internado	246800
Alta	4000
Alta	502
Alta	70
Não Internado	203
Não Internado	5
Não Internado	6
Não Internado	14
Óbito	400
Óbito	29

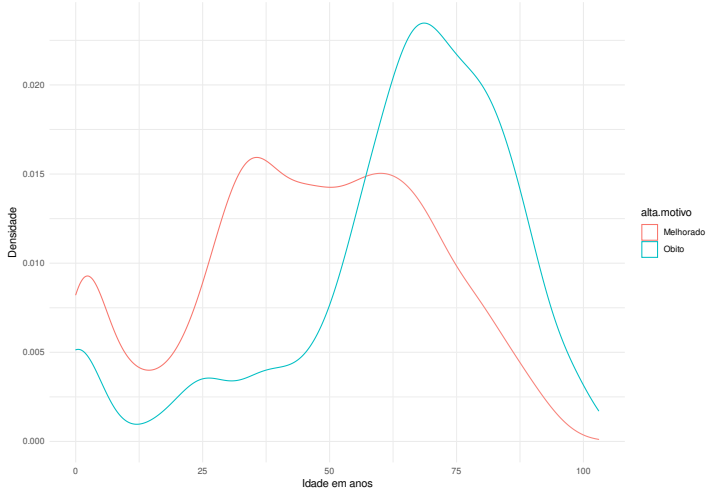
Exemplo: Desfecho de pacientes internados em um hospital

Desfecho	Quantidade
Internado	246800
Alta	4572
Não Internado	228
Óbito	429

Exemplo: Desfecho de pacientes internados em um hospital

Desfecho	Quantidade
Melhorado	4572
Óbito	429

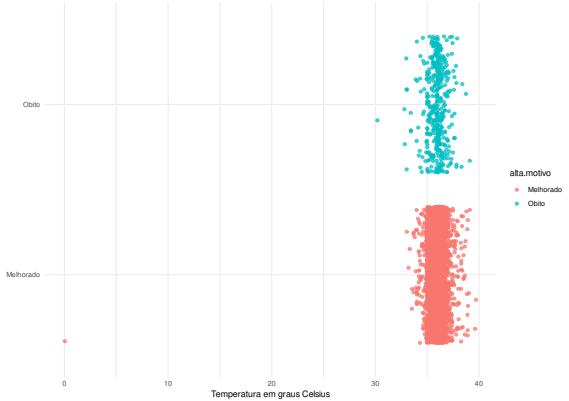
Exemplo: Desfecho de pacientes internados em um hospital



Exemplo: Desfecho de pacientes internados em um hospital

	<=55 anos	>55 anos
Melhorado	2830	1742
Obito	101	328

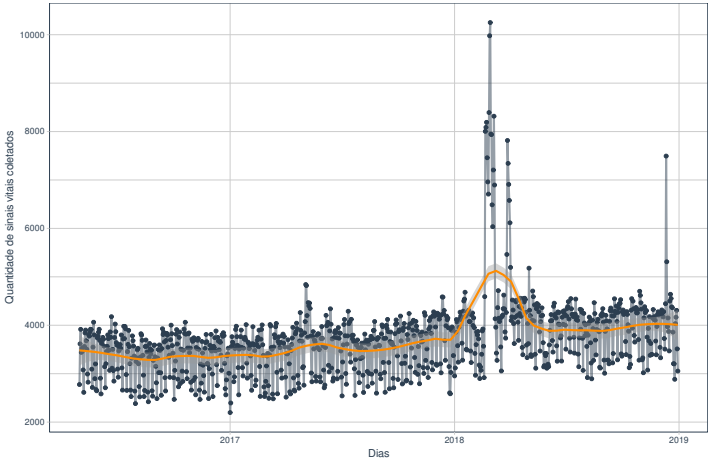
Exemplo: Desfecho de pacientes internados em um hospital



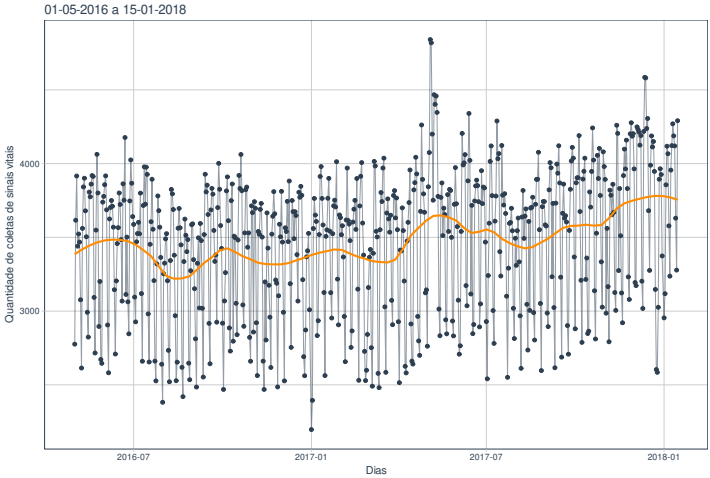
Exemplo: Desfecho de pacientes internados em um hospital


```
Melhorado      Obito
0.3526037 0.8087152
```

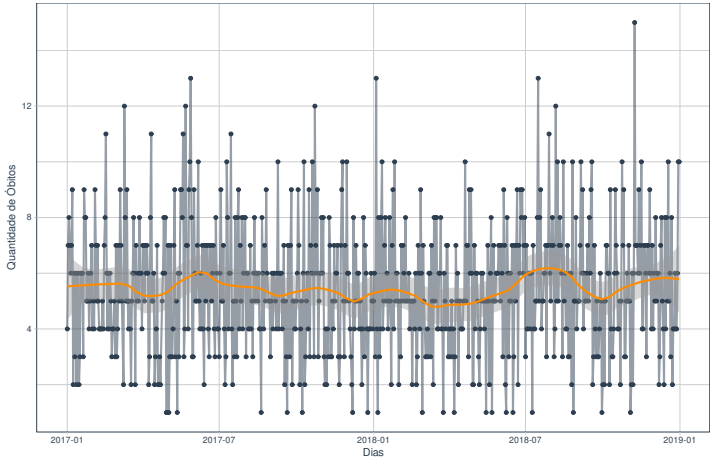
Exemplo: Sinais vitais coletados de pacientes internados em um hospital



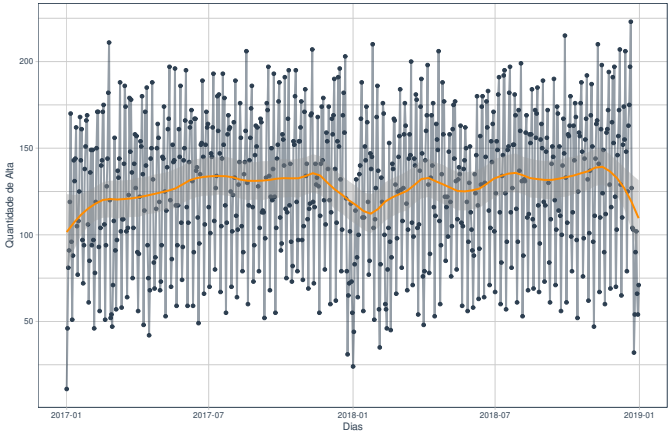
Exemplo: Sinais vitais coletados de pacientes internados em um hospital



Exemplo: Óbitos de pacientes internados em um hospital



Exemplo: Altas de pacientes internados em um hospital

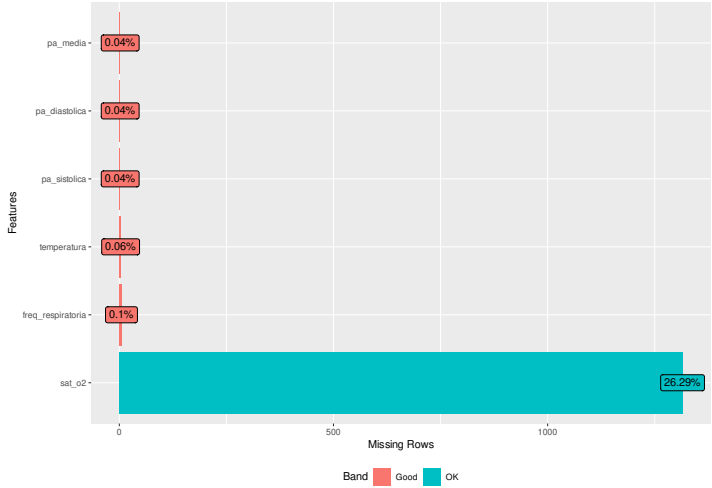


domingo	quarta	quinta	sábado	segunda	sexta	terça
61	144	148	107	97	174	135

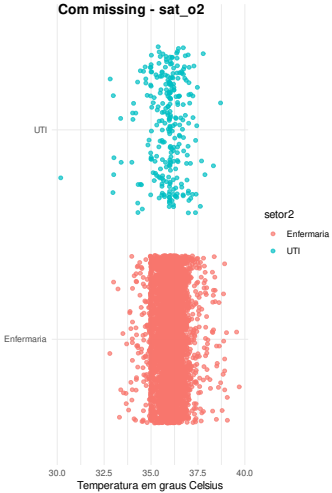
Dados faltantes - **missing data**

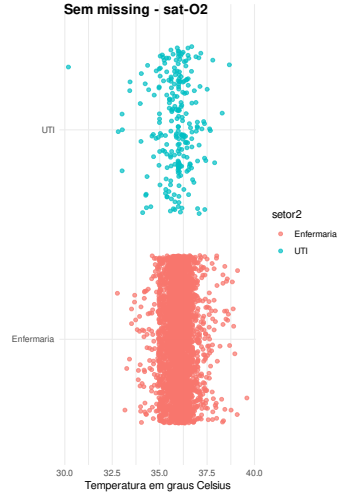
- ***Missing completely at random (MCAR):*** quando a probabilidade dos dados faltantes é a mesma entre as observações. Por exemplo, dados perdidos por um backup incorreto;
- ***Missing at random (MAR):*** quando os dados faltantes variam de acordo com outras variáveis. Por exemplo, missing em uma variável idade pode ser diferente entre mulheres e homens;
- ***Missing not at random (MNAR):*** quando a probabilidade de missing está relacionada ao missing. Por exemplo, dependendo da renda do cliente, é mais provável que ele não responda sobre a renda.

Exemplo - dados faltantes



Exemplo - dados faltantes





Exemplo - dados faltantes

- Distribuição do desfecho na **UTI**

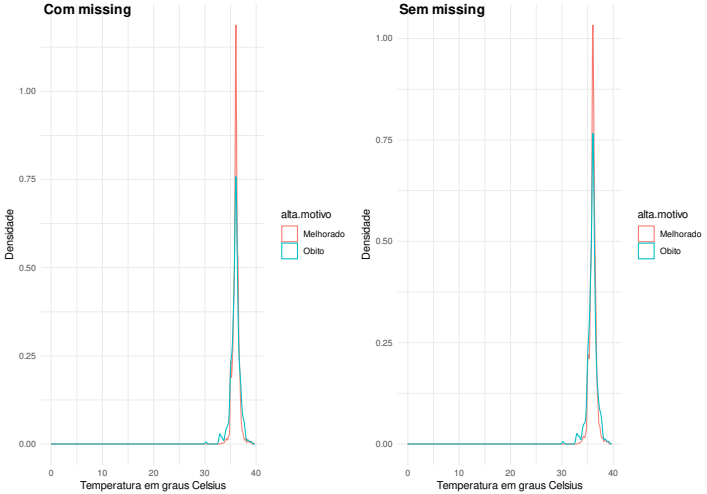
Com missing - sat_o2

```
Melhorado      Obito
   0.0438     0.9562
```

Sem missing - sat_o2

```
Melhorado      Obito
   0.0412     0.9588
```

Exemplo - dados faltantes

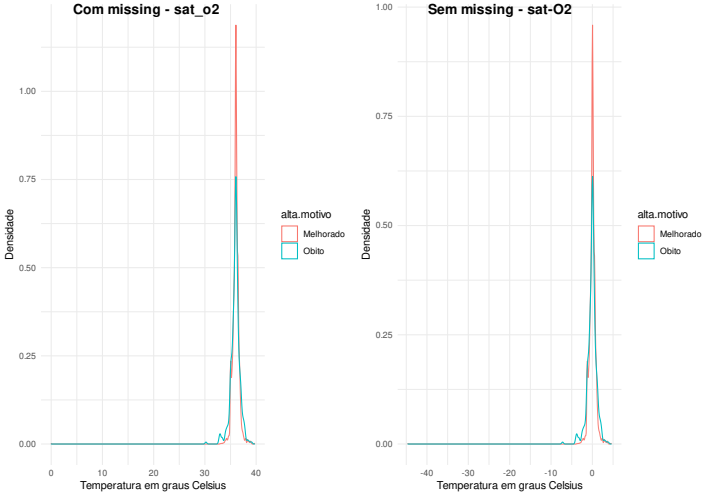


Imputação

- Algoritmo **KNN** *(k-nearest neighbors)*

1. Recebe um dado não classificado;
2. Define o tamanho da vizinhança (K);
3. Mede a distância (Euclidiana) do novo; dado com todos os outros dados que já estão classificados;
4. Obtém a menor ou menores distâncias;
5. Verifica o valor de cada da um dos dados que tiveram a menor distância;
6. Toma como resultado o valor que mais apareceu dentre os dados que tiveram as menores distâncias.

Imputação


Com missing - sat_o2
Densidade
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
Temperatura em graus Celsius
alta.motivo
Melhorado
Obito
Sem missing - sat-O2
Densidade
1.00
0.75
0.50
0.25
0.00
-40
-30
-20
-10
0
Temperatura em graus Celsius
alta.motivo
Melhorado
Obito

Distribuições de probabilidade

- Há diversos modelos probabilísticos
- Distribuições de probabilidade de V.A. (discretas ou contínuas)
- Descrição das probabilidades associadas com os possíveis valores de X
  - **Variáveis discretas** $\Rightarrow$ suporte em um conjunto de valores enumeráveis (finitos ou infinitos)
  - **Variáveis contínuas** $\Rightarrow$ suporte em um conjunto não enumerável de valores

Distribuição de probabilidade - variável aleatória contínua

Definição: função densidade de probabilidade

A função densidade de probabilidade $f(x)$ de uma V.A. contínua, é uma função definida em um intervalo de valores na reta dos reias, ou seja,

$$P[a \leq X \leq b] = \int_b^a f_x(x).$$

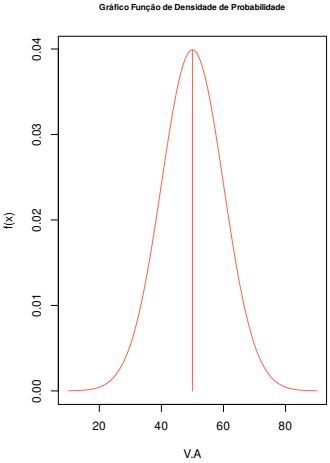
i. A área total da abaixo da curva de $f(x)$ é 1.

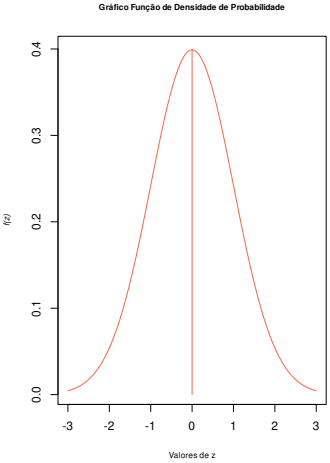
$$\int_{-\infty}^{\infty} f_x(x) = 1$$

ii. A função $f(x)$ é positiva ou zero.

$$f_x(x) \geq 0$$

Padronização





Modelo Gaussiano(ou Normal)

Definição:

Uma V.A. contínua é dita ter distribuição gaussiana com parâmetros μ e σ^2 se sua f.d.p. for:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty$$

Notação: $X \sim \mathrm{N}(\mu, \sigma^2)$

Normal padrão

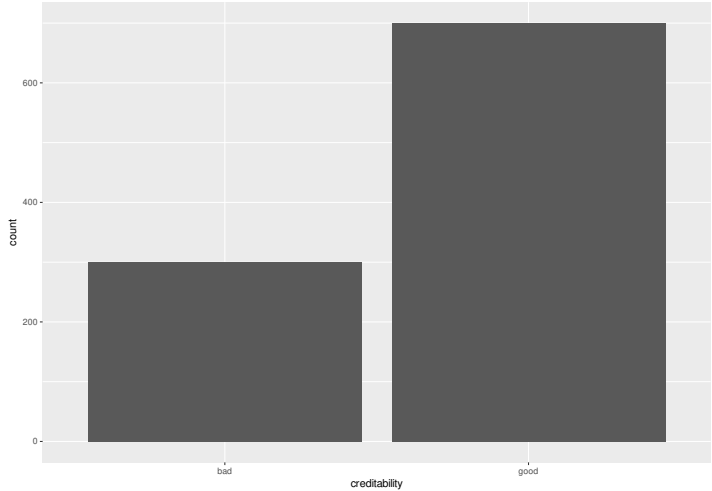
Definição:

Uma V.A. contínua que segue uma distribuição normal com parâmetros $\mu = 0$ e $\sigma^2 = 1$, é conhecida com uma distribuição normal padrão.
Notação: $Z \sim \mathrm{N}(\mu = 0, \sigma^2 = 1)$
e sua f.d.p. é:

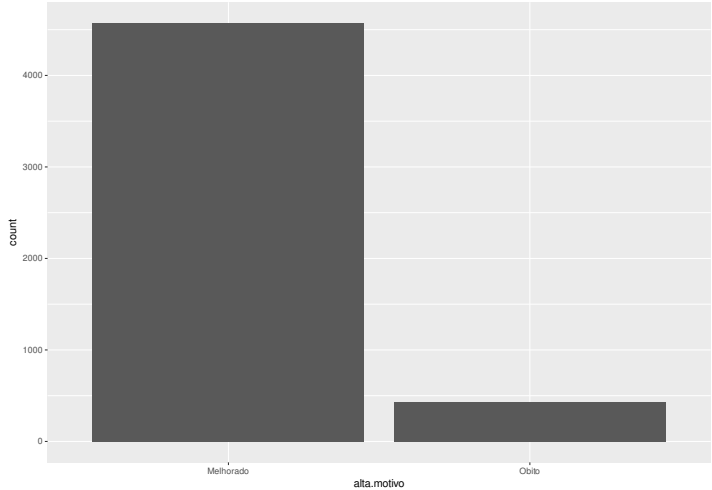
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(z)^2\right], \quad -\infty < z < \infty$$

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

Dados desbalanceados - Exemplo 1



Dados desbalanceados - Exemplo 2



Probabilidade condicional

Intuitivamente, a probabilidade de ocorrer um evento pode muito bem ser influenciada pelo ocorrência ou não de outros eventos

Exemplo:

- Probabilidade de que um aluno não comparecer a uma aula será influenciado:
  - esteja chovendo ou não,
  - trote dos calouros foi ou não realizado na noite anterior.
  - Internet falhou ou não
- Esta é a probabilidade condicional

Probabilidade condicional

Intuitivamente, a probabilidade de ocorrer um evento pode muito bem ser influenciada pelo ocorrência ou não de outros eventos

Exemplo:

- Probabilidade de que um aluno não comparecer a uma aula será influenciado:
  - esteja chovendo ou não,
  - trote dos calouros foi ou não realizado na noite anterior.
  - Internet falhou ou não
- Esta é a probabilidade condicional

Probabilidade condicional

Definição

- Dados dois eventos A e B, a probabilidade condicional de A ocorrer, dado que ocorreu B é representado por $P(A|B)$ e dada por

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{para} \quad P(B) > 0.$$

Regra do produto

- Da definição de probabilidade condicional

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

temos que

$$P(A \cap B) = P(A|B) \cdot P(B).$$

- Caso $P(B) = 0$, definimos $P(A|B) = P(A)$.

Independência de eventos

Os eventos A e B são **eventos independentes** se a ocorrência de B não altera a probabilidade de ocorrência de A, ou seja, eventos A e B são independentes se

$$P(A|B) = P(A) \quad \text{e também que} \quad P(B|A) = P(B).$$

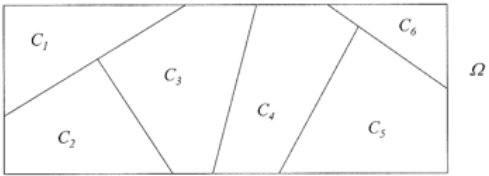
Com isso, e a regra do produto, temos que

$$P(A \cap B) = P(B) \cdot P(A|B) = P(B) \cdot P(A).$$
$$P(A \cap B) = P(A) \cdot P(B|A) = P(A) \cdot P(B).$$

Partição do espaço amostral

Dizemos que os eventos $C_1, C_2, ..., C_k$ formam uma **partição** do espaço amostral, se eles não tem interseção entre si, e se sua união é igual ao espaço amostral. Isto é,

$$C_i \cap C_j = \emptyset \quad \text{para} \quad i \neq j \quad \text{e} \quad \bigcup_{i=1}^{k} C_i = \Omega.$$



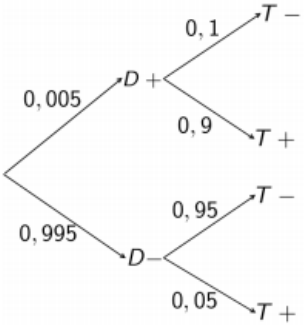
Teorema de Bayes

Suponha que os eventos $C_1, C_2, \ldots, C_k$ formem uma partição de Ω e que suas probabilidades sejam conhecidas. Suponha, ainda, que para um evento A, se conheçam as probabilidades $P(A|C_i)$ para todo $i = 1, 2, \ldots, k$. Então, para qualquer j,

$$P(C_j|A) = \frac{P(C_j)P(A|C_j)}{\sum_{i=1}^{k} P(C_i)P(A|C_i)}, \quad j = 1, 2, \ldots, k.$$

Exemplo - Regra de Bayes

- Considere um teste de triagem de rotina para uma doença



- Teste altamente preciso
  - $P(FP) = P(T_+|D_-) = 0.05$
  - $P(FN) = P(T_-|D_+) = 0.10$

Exemplo - Regra de Bayes

- Qual a probabilidade de ter a doença, dado que o teste foi positivo?
- $P(D_+|T_+) = \frac{P(D_+) * P(T_+|D_+)}{P(T_+)}$

- Calculamos o denominador utilizando a lei da probabilidade total:
- $P(T_+) = P(D_+) * P(T_+|D_+) + P(D_-) * P(T_+|D_-) = 0.005 * .9 + 0.995 * .05 = 0.05425$

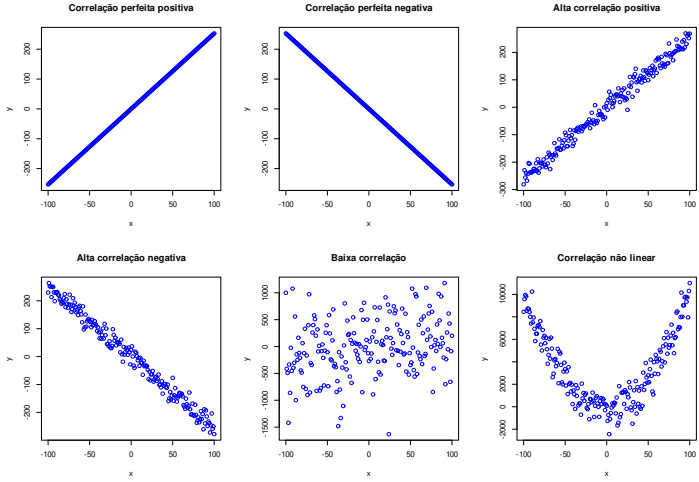
Assim,

- $P(D_+|T_+) = \frac{P(D_+) * P(T_+|D_+)}{P(T_+)} = \frac{0.005 * 0.9}{0.05425} = .083$

Dados desbalanceados

- **SMOTE** (*Synthetic Minority Over-sampling Technique*): reamostrar o conjunto de dados original por superamostragem da classe minoritária.
- Utilizar uma amostra balanceda excluindo registros da classe majoritária.
- Definir um ponto de corte diferente do *default*.

Correlação


Correlação perfeita positiva
Correlação perfeita negativa
Alta correlação positiva
Alta correlação negativa
Baixa correlação
Correlação não linear
x
y

Correlação

Coeficientes de correlação linear

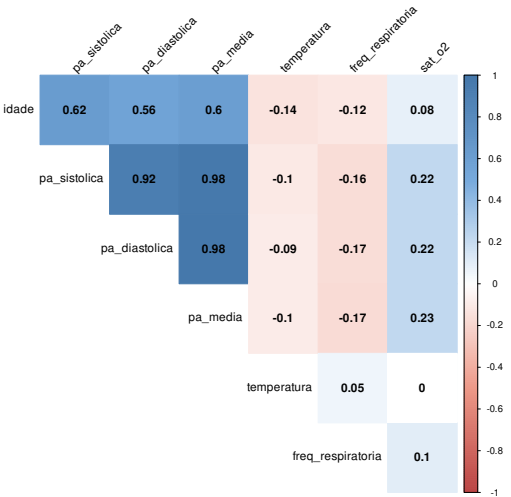
- ***Spearman***

$$r = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

- ***Pearson***

$$\rho(X_1, X_2) = \frac{cov(X_1, X_2)}{\sigma_{x1}\sigma_{x2}}$$

Matriz de correlação



Desafio2

Com os dados coletados no desafio anterior, faça:

- Engenharia de características
  1. Faça uma análise exploratória de seus dados. Com gráficos e medidas descritivas;
  2. Crie novas características com base em outras já presente no conjunto de dados;
  3. Se houver dados faltantes, identifique-os e use dois tipos de imputação e compare os resultados;
  4. Se houver dados desbalanceados em sua label(Y), utilize a ténica *SMOTE* ou redução da classe majoritária;
  5. Faça uma matriz de correlação linear, interprete e decida se irá excluir ou não alguma característica.

Desafio2

Com os dados coletados no desafio anterior, faça:

- Engenharia de características
  1. Faça uma análise exploratória de seus dados. Com gráficos e medidas descritivas;
  2. Crie novas características com base em outras já presente no conjunto de dados;
  3. Se houver dados faltantes, identifique-os e use dois tipos de imputação e compare os resultados;
  4. Se houver dados desbalanceados em sua label(Y), utilize a ténica *SMOTE* ou redução da classe majoritária;
  5. Faça uma matriz de correlação linear, interprete e decida se irá excluir ou não alguma característica.

Desafio2

Com os dados coletados no desafio anterior, faça:

- Engenharia de características
  1. Faça uma análise exploratória de seus dados. Com gráficos e medidas descritivas;
  2. Crie novas características com base em outras já presente no conjunto de dados;
  3. Se houver dados faltantes, identifique-os e use dois tipos de imputação e compare os resultados;
  4. Se houver dados desbalanceados em sua label(Y), utilize a ténica *SMOTE* ou redução da classe majoritária;
  5. Faça uma matriz de correlação linear, interprete e decida se irá excluir ou não alguma característica.

Desafio2

Com os dados coletados no desafio anterior, faça:

- Engenharia de características
  1. Faça uma análise exploratória de seus dados. Com gráficos e medidas descritivas;
  2. Crie novas características com base em outras já presente no conjunto de dados;
  3. Se houver dados faltantes, identifique-os e use dois tipos de imputação e compare os resultados;
  4. Se houver dados desbalanceados em sua label(Y), utilize a ténica *SMOTE* ou redução da classe majoritária;
  5. Faça uma matriz de correlação linear, interprete e decida se irá excluir ou não alguma característica.

Desafio2

Com os dados coletados no desafio anterior, faça:

- Engenharia de características
  1. Faça uma análise exploratória de seus dados. Com gráficos e medidas descritivas;
  2. Crie novas características com base em outras já presente no conjunto de dados;
  3. Se houver dados faltantes, identifique-os e use dois tipos de imputação e compare os resultados;
  4. Se houver dados desbalanceados em sua label(Y), utilize a ténica *SMOTE* ou redução da classe majoritária;
  5. Faça uma matriz de correlação linear, interprete e decida se irá excluir ou não alguma característica.

Referências

1. Hastie, T., Tibshirani, R. e Friedman, J., **The Elements of Statistical Learning**, 2009.
2. James, G., Witten, D., Hastie, T. e Tibshirani, **An Introduction to Statistical Learning**, 2013.
3. L. Breiman. Statistical modeling: **The two cultures. Statistical Science**, 16(3):199-231, 2001.
4. Lantz, B., **Machine Learning with R**, Packt Publishing, 2013.
5. Tan, Steinbach, and Kumar, **Introduction to Data Mining**, Addison-Wesley, 2005.
6. Mitchell, T. M. (1997). **Machine Learning**. McGraw-Hill.
7. Wickham, H.; Grolemund, G. **R for data science: import, tidy, transform, visualize, and model data**. " O'Reilly Media, Inc.", 2016.