



Insper

Ciência dos dados

Teorema de Bayes e Classificador

Aula de hoje

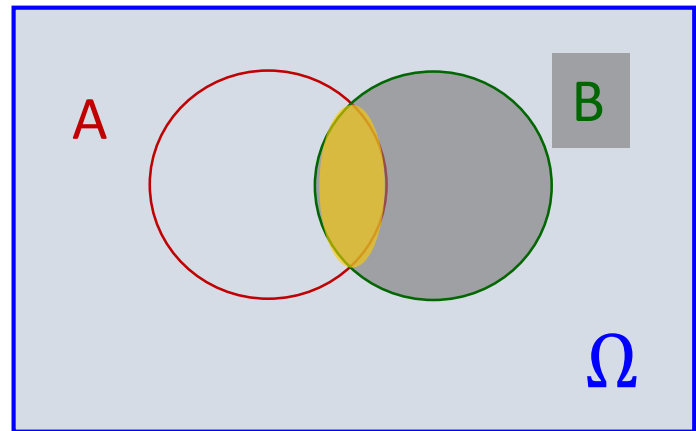
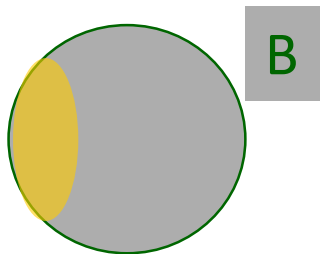
Ao final desta aula, o aluno deve ser capaz de:

- Entender conceitos básicos do Teorema de Bayes
- Compreender e implementar um classificador Naive Bayes.

Probabilidade Condicional

$$P(A|\mathbf{B})$$

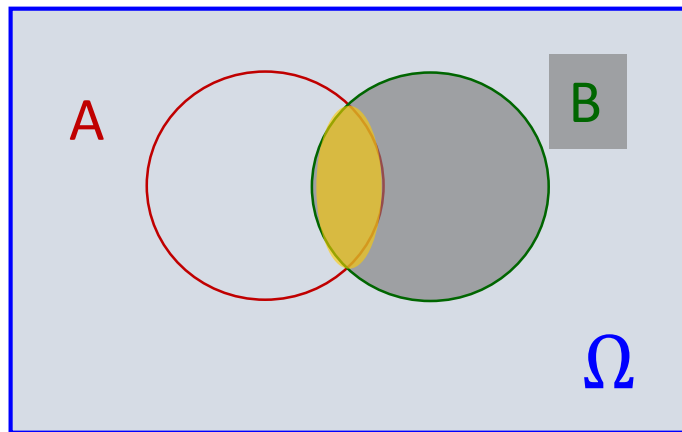
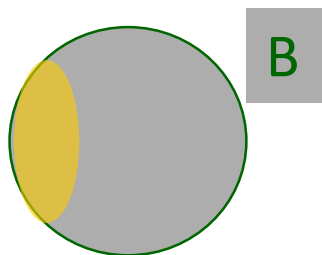
Novo espaço amostral:



Probabilidade Condicional

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

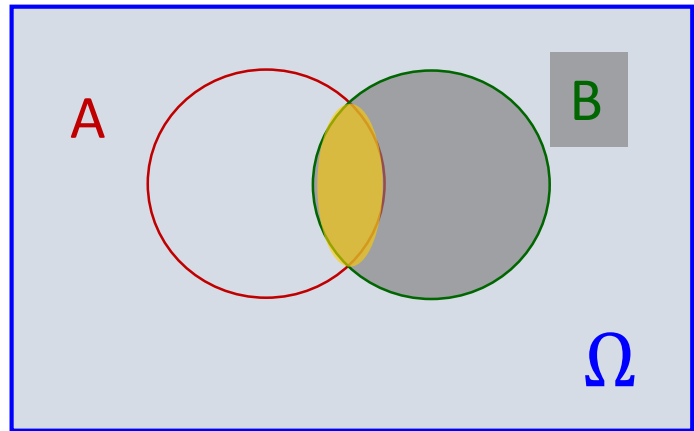
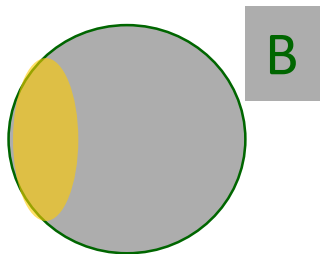
Novo espaço amostral:



Intersecção

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

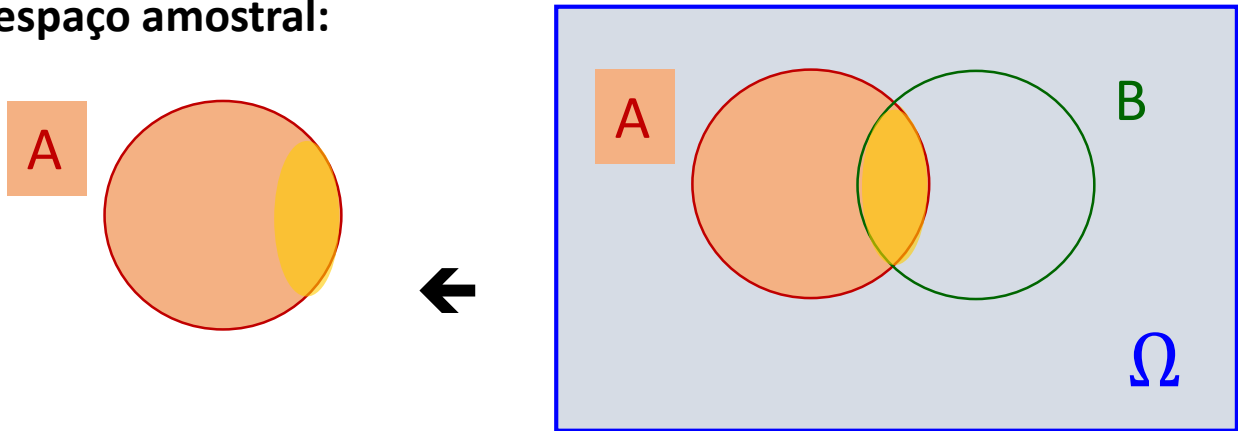
Novo espaço amostral:



Probabilidade Condicional

$$P(B|A)$$

Novo espaço amostral:

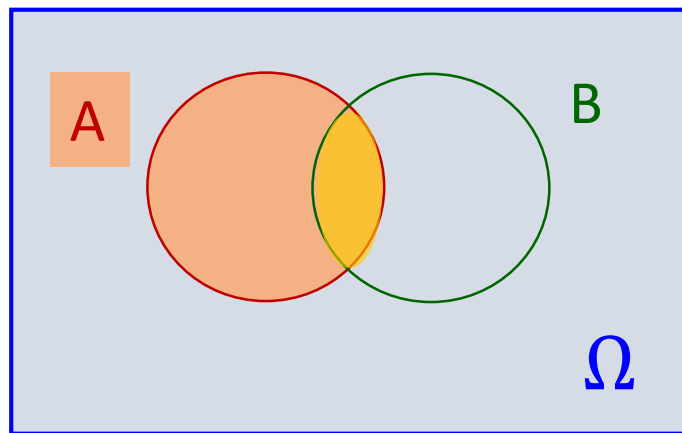
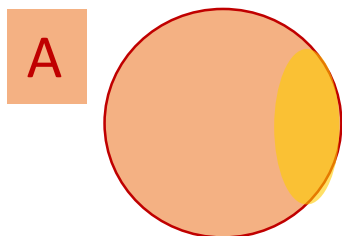


Probabilidade Condicional

7

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

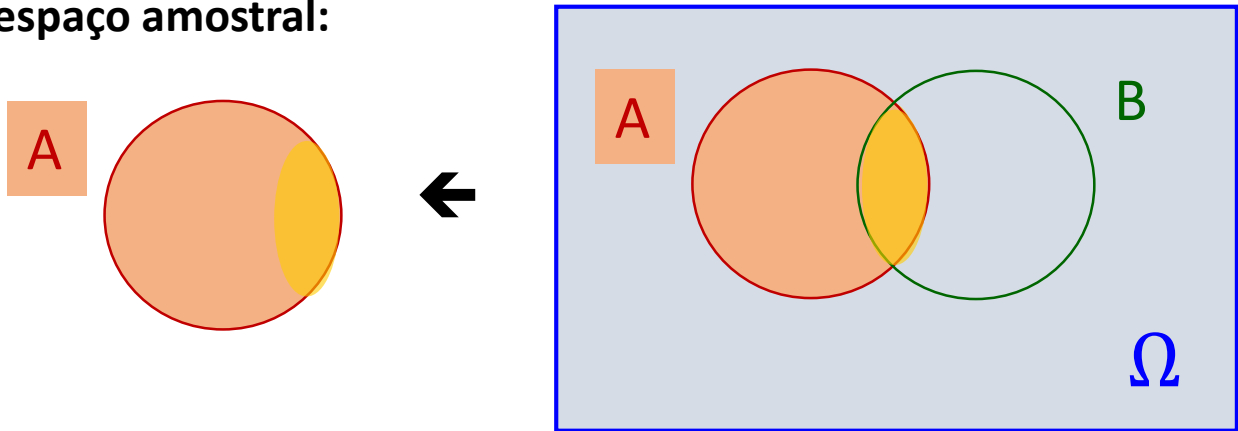
Novo espaço amostral:



Intersecção

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

Novo espaço amostral:



Teorema de Bayes

$$\text{Se } P(A | B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A | B)P(B) \quad (I)$$

$$\text{E } P(B | A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B | A)P(A) \quad (II)$$

Igualando (I) e (II), então:

$$P(A | B)P(B) = P(B | A)P(A)$$

O que nos traz ao **Teorema de Bayes**:

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

Exemplo: Gostar de assistir filmes x Sexo.

A preferência por assistir filmes depende de Sexo?

Frequências Relativas

Assiste Filmes?	Sexo		Total
	Masculino	Feminino	
Sim	42%	18%	60%
Não	28%	12%	40%
Total	70%	30%	100%

S: gosta de assistir filmes

H: homem

M: mulher

Entre os homens, qual a probabilidade de gostar de assistir filmes? E entre as mulheres?

$$P(S | H) = \frac{P(S \cap H)}{P(H)} = \frac{0,42}{0,70} = 0,60$$

$$P(S | M) = \frac{P(S \cap M)}{P(M)} = \frac{0,18}{0,30} = 0,60$$

Notamos que

$$P(S | H) = P(S) \text{ e } P(S | M) = P(S)$$

Logo, a preferência por assistir filmes independe do sexo.

Independência

Os eventos A e B são independentes quando o fato de ter conhecimento sobre a ocorrência de A **não altera** a expectativa sobre a probabilidade de ocorrência do evento B.

Atenção: não confundir eventos **independentes** com eventos **disjuntos** (mutuamente exclusivos).

$P(A \cap B) = P(A)P(B) \Rightarrow$ **Independentes**

$P(A \cap B) = 0 \Rightarrow$ **Disjuntos**

Exercício 1 da Aula 06

13

A probabilidade de que o preço dos combustíveis aumente no mês vindouro é estimada em 0,4.

Se isto ocorrer, a probabilidade de que os preços dos transportes coletivos também aumentem é de 0,5; caso contrário, esta probabilidade é de 0,1.

Se naquele mês o preço das passagens de fato subirem, qual a probabilidade de os preços dos combustíveis não terem sofrido majoração?

Resposta: 0,231

Extra - Exercício 1

14

Analizando-se o histórico de acidentes provocados por clientes de uma seguradora, sabe-se que 10% dos contratos resultam em sinistros. A probabilidade de um homem causar um sinistro é 12%, enquanto que a probabilidade de uma mulher causar um sinistro é de 0,06.

- a. Sorteado um contrato ao acaso, qual é a probabilidade de ser de um homem? 0,667
- b. Sorteado um contrato no qual ocorreu um sinistro, qual é a probabilidade de ter sido causado por uma mulher? 0,20

Extra - Exercício 2

15

Em uma região, 3 indústrias dividem o mercado de celulares. A empresa A vende o dobro que a empresa B. A empresa A vende a mesma quantidade de celulares que C. De todos os celulares vendidos, 10% dos celulares foram comercializados pela empresa A e são pós-pagos. Sabe-se que 15% dos celulares comercializados pela empresa B são pós-pagos. Além disso, sabe-se que do total de celulares pós-pagos 25% foram comercializados pela empresa C.

a) Qual é a probabilidade de um celular não ter sido comercializado por C? Resposta: $\frac{3}{5}$

b) Se for escolhido um celular ao acaso, qual é a probabilidade de ser pós-pago? Resposta: 0,173

c) Qual é a probabilidade de um celular ser pós-pago e não ser da indústria B? Resposta: 0,143

Extra - Exercício 2

16

Em uma região, 3 indústrias dividem o mercado de celulares. A empresa A vende o dobro que a empresa B. A empresa A vende a mesma quantidade de celulares que C. De todos os celulares vendidos, 10% dos celulares foram comercializados pela empresa A e são pós-pagos. Sabe-se que 15% dos celulares comercializados pela empresa B são pós-pagos. Além disso, sabe-se que do total de celulares pós-pagos 25% foram comercializados pela empresa C.

Eventos:

A: empresa A

B: empresa B

C: empresa C

P: plano pós-pago

Extra - Exercício 3

17

Uma fábrica de parafusos tem exatamente três máquinas trabalhando na linha de produção, aqui, nomeadas de: M1, M2 e M3. A máquina M2 produz 28% do total de parafusos e a M3 é responsável por 39% da produção total. Para o setor de Controle de Qualidade, cada parafuso (produzido por uma das três máquinas) pode ser classificado como: perfeito, defeituoso recuperável ou defeituoso descartável.

Do total de parafusos produzidos, 81% são classificados como perfeitos.

Da produção total, 2% são produzidos pela máquina M2 e classificados como defeituosos descartáveis.

Dos parafusos produzidos pela máquina M1, 75,8% são classificados como perfeitos. Da produção da máquina M2, 92,8% são classificados como perfeitos.

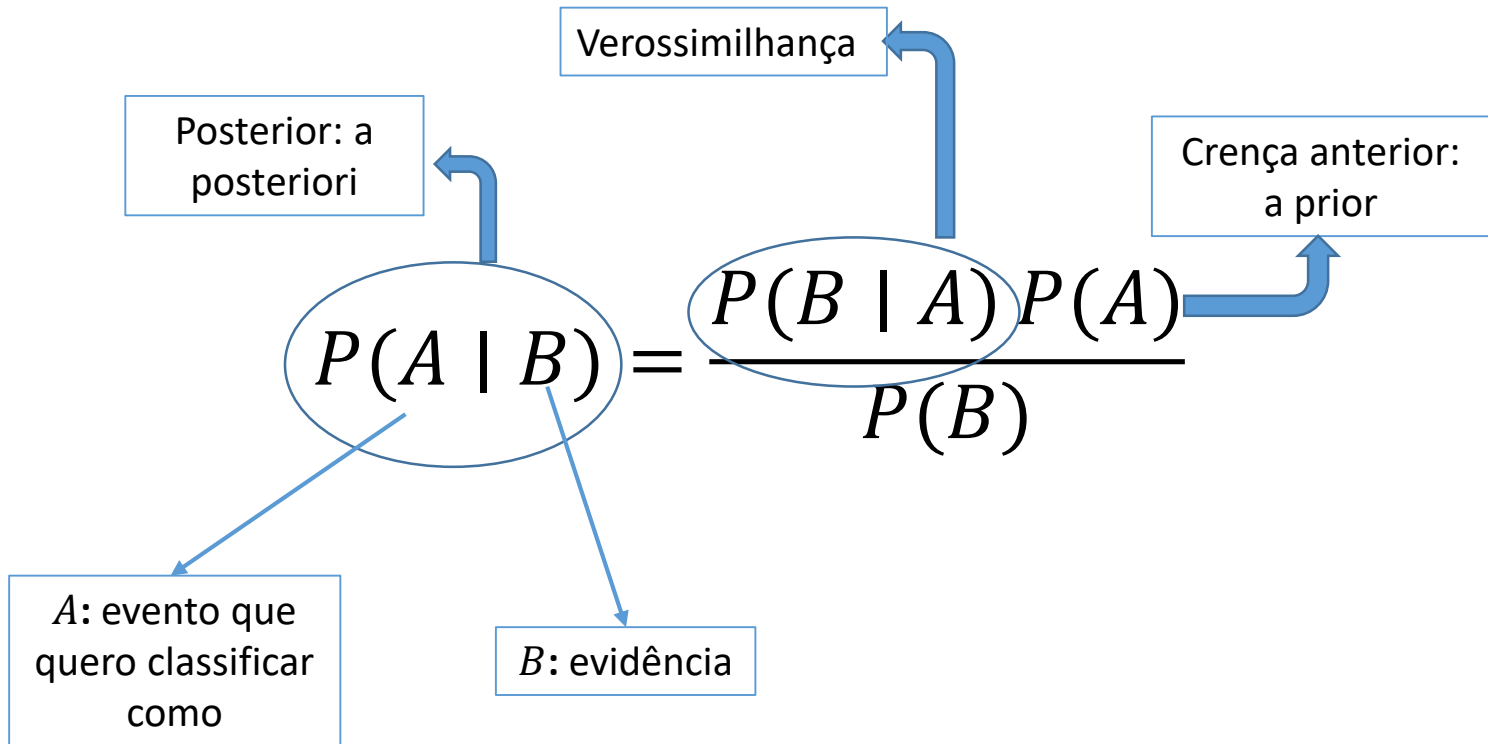
Apenas considerando o que é produzido pela máquina M3, 17,9% são classificados como defeituosos recuperáveis. Entretanto, de todos os parafusos defeituosos recuperáveis, 41,7% são produzidos pela máquina M1.

Extra - Exercício 3 - continuação

Responda:

- a) Calcule a probabilidade de um parafuso ser produzido pela máquina M2 e classificado como perfeito. R.: 0,260
- b) Considerando que um parafuso é classificado como perfeito, qual a probabilidade dele ter sido produzido pela máquina M1?
R.: 0,309
- c) Qual a probabilidade de um parafuso ter sido produzido pela máquina M3 dado que esse foi classificado como perfeito?
R.: 0,370
- d) Calcule a probabilidade de um parafuso ser classificado como defeituoso recuperável.
R.: 0,120

Teorema de Bayes



Titanic

Entre as muitas competições disponíveis na plataforma Kaggle, encontra-se a de *Machine Learning* do Titanic conforme apresentado neste [link](#).

A base de dados apresentada no arquivo `titanic.xlsx` contém informação de 891 passageiros para os quais foram mensurados as seguintes informações (as categorias serão mantidas em inglês assim como os nomes das variáveis):

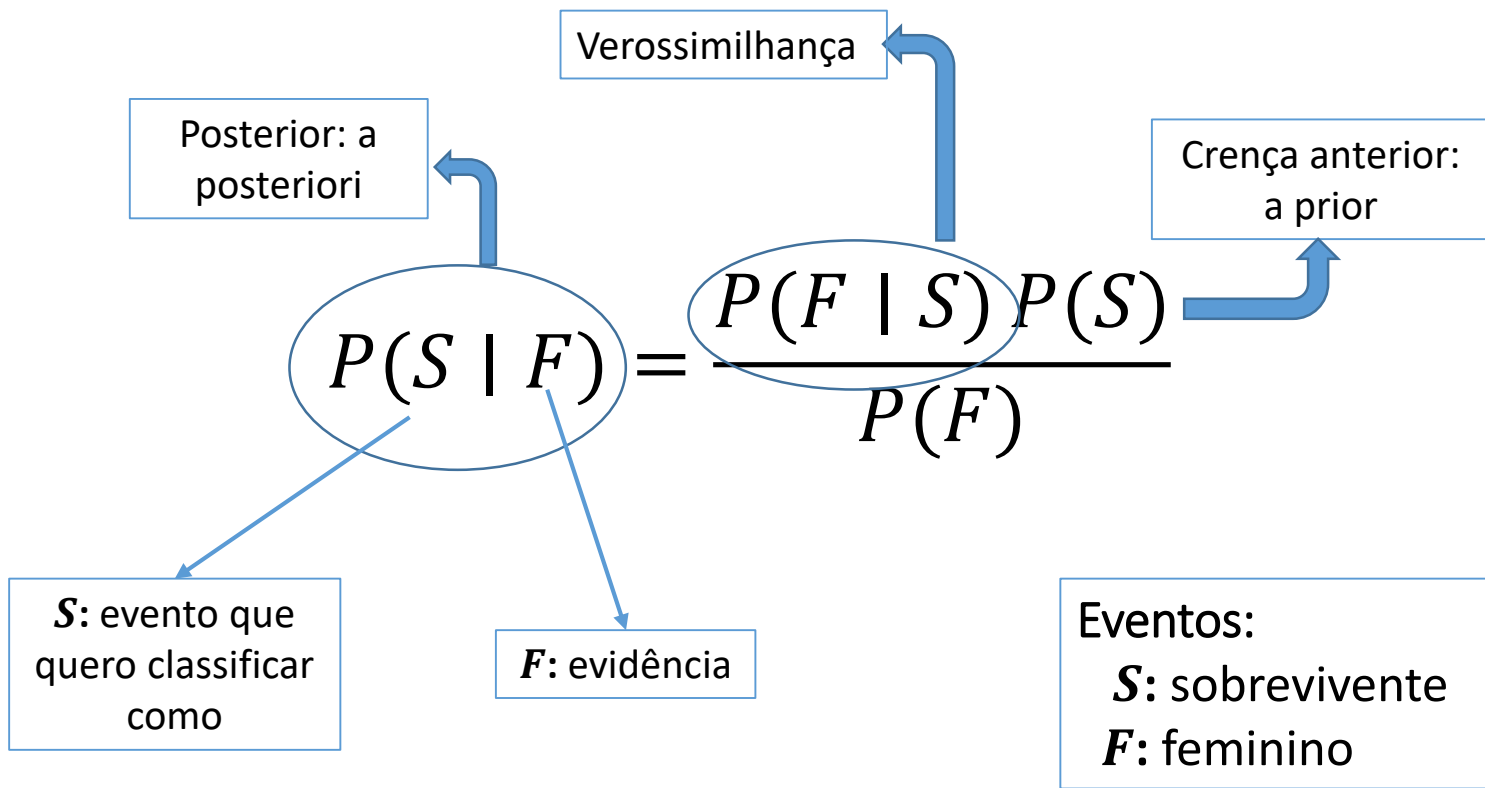
- `Survived` : indica se passageiro sobreviveu ou não.
 - 0 = No,
 - 1 = Yes.
- `Pclass` : indica a classe da passagem (*ticket*).
 - 1 = 1st,
 - 2 = 2nd,
 - 3 = 3rd.
- `Sex` : sexo do passageiro.
 - female,
 - male.

Objetivo

Imagine que você com as suas características estivesse no Titanic no dia do naufrágio. Será que você teria sido um sobrevivente ou não??

Assim, o objetivo dessa atividade é classificar um "novo" passageiro do Titanic como Sobrevivente ou não Sobrevivente levando em consideração (ou seja, aprendendo) o perfil dos sobreviventes e não sobreviventes que de fato estavam no Titanic.

Teorema de Bayes - Titanic



- Download do notebook pelo Github
- Download dos slides pelo Blackboard
- Fazer individual e discutir em grupo

- Usar arquivo:

Aula07_Atividade_Titanic....ipynb