
FIAP

AULA 5

"Capacitar o aluno na construção de protótipos de Data Marts e Data Warehouse, utilizando ETL e a evolução para Data Ingestion."

"Tornar o aluno capacitado a compreender e construir modelos estatísticos/analíticos básicos de classificação e preditivos utilizados em Data Science/Machine Learning."

Enterprise Analytics e Data Warehousing

Alcides C. Araújo

Problema

- Uma empresa bagunçada possui 4 sistemas para gerenciar sua base de clientes digitais.

- Para piorar a situação, os sistemas não conversavam de forma sincronizada.

- Seus clientes precisam ser classificados como Opt-In (deseja receber emails), Opt-out (não deseja receber emails) e Drop-out (desejava receber e-mails, mas pediu para sair da lista de subscritos)

Problema

- Desenho dos sistemas:

Gerenciamento de entrada 1	Gerenciamento de entrada 2	Gerenciamento de saída 1	Gerenciamento de saída 2
Opt-In	Opt-In	Unsubscribe	Unsubscribe
Opt-Out	Opt-Out	-	-

- Como os sistemas não conversavam, um cliente poderia estar presente em uma das bases ou em todas ao mesmo tempo.

Problema

- Principais perguntas da área:
- Qual a probabilidade de um novo cliente ser Opt-In?
- Qual a probabilidade de um novo cliente ser Opt-In dado que está como Opt-Out?
- Desafio implícito: quantas regras teremos que criar para organizar esses bancos e termos uma classificação única (Opt-In, Opt-Out e Drop-Out) para cada cliente?

Conceitos iniciais

- Probabilidade

Conceito

- Medida numérica sobre a possibilidade de um determinado evento ocorrer

- A forma simples de representar seria:

$$P(x) = \frac{\#\ de\ eventos\ desejados}{\#\ de\ eventos\ totais}$$

Conceito

- O conceito de probabilidade está relacionado aos seguintes temas:

- Princípios de contagens de eventos
- Permutação
- Combinação
- Condicionalidade

- Iremos trabalhar nestes temas para resolver o problema inicial.

Contagens de eventos

- A contagem de eventos é a atividade de obter todas a possibilidades dentro de um conjunto ações realizadas.

- Um evento é uma ocorrência, podendo ocorrer ou não ocorrer.

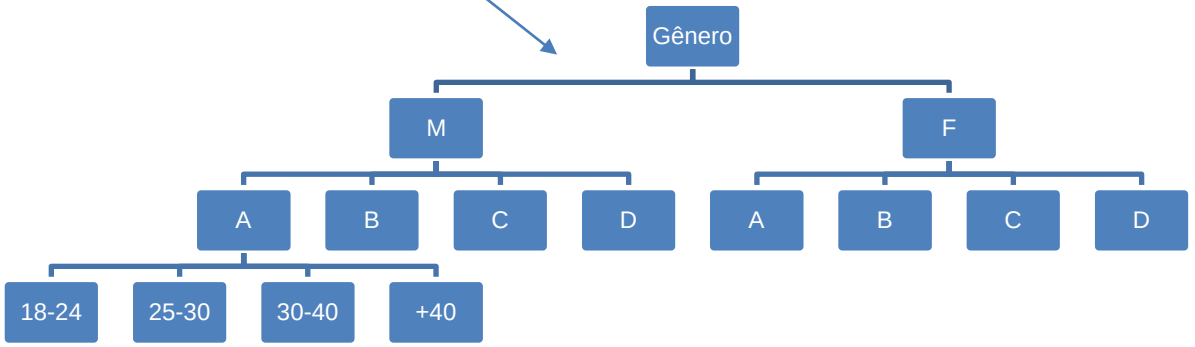
- Dado uma sequência de eventos, estes podem ser contados de forma discreta (dias, número de vendas, número de acidentes) ou contínua (taxa de crescimento de uma empresa, renda de uma família)

Contagens de eventos

- No exemplo temos um banco de dados e ao lado a contagem dos seus eventos:

Faixa Etária	Renda	Gênero
18-24	A	Masculino (M)
25-30	B	Feminino(F)
30-40	C	
+40	D	

Exemplo de como contar o número de eventos possíveis



Contagens de eventos

- A forma de contar o número de eventos irá depender de como podemos observar o problema.

- Ou seja, no momento de ordenar as ações a serem realizadas, a ordem de como ela é construída irá importar?

- Neste momento, precisaremos entender os conceitos de permutação e combinação.

Permutação

- Quando preciso arranjar eventos, sendo que a **ordem importa**, temos um problema de **permutação**.

Temos 6 maneiras de arranjar as 3 bolas nas cores diferentes.

Apesar das cores serem as mesmas, o fato de estarem arranjadas em ordens diferentes temos várias permutações, no caso 6.

Combinação

- Quando preciso arranjar eventos, **sem considerar a ordem**, temos um problema de **combinação**.

Temos 4 pessoas para arranjar em grupos de 2 pessoas.

A ordem como iremos dividir as pessoas nesses grupos não importa. Sendo ao todo 6 combinações.

Uma dupla com a pessoa de camisa vermelha e a pessoa de camisa verde será a mesma, não importando se eu escolhi o verde ou o vermelho primeiro.

Fórmulas

- Permutações

- $P_k^n = \frac{n!}{(n-k)!}$, sem repetição
- $P_k^n = n^k$, com repetição

- Combinações

- $C_k^n = \frac{n!}{k!(n-k)!}$, sem repetição
- $C_k^n = \frac{(n+k-1)!}{k!(n-1)!}$, com repetição

Permutação ou Combinação?

- Precisamos escolher uma senha contendo **6** caracteres, sendo que estes caracteres somente podem ser **números** e **não podem ser repetidos**.

  - Números podem escolhidos entre 0 – 9. Temos 10 maneiras para arranjar.
  - Temos que escolher 6 números
  - A ordem dos números importa? Pense comigo, uma senha 123456 é a mesma que 654321?

  - Este é um problema de permutação sem repetição.
  - $P_6^{10} = \frac{10!}{(10-6)!} = 151.200$ possíveis permutações diferentes de senhas

Permutação ou Combinação?

- Na Megasena precisamos escolher **6** números dentro de **60** possibilidades, quando um número é sorteado, ela **não pode aparecer novamente**.

  - Números podem escolhidos entre 1 – 60.
  - Temos que escolher 6 números
  - A ordem do jogo importa? Pense comigo, uma aposta (50, 20, 31, 37, 10, 45) é a mesma que (20, 31, 50, 10, 45, 37)?

  - Este é um problema de combinação sem repetição.
  - $C_6^{60} = \frac{60!}{6!(60-6)!}$ possíveis combinações.
    Aproximadamente 50 milhões de combinações.

Condicionalidade

- Observe a seguinte tabela:

	Data Scientist	*Data Engineer*	Total
M	27	9	36
F	10	14	24
Total	37	23	**60**

- Qual a probabilidade de, numa amostra, ter 1 Mulher (F)?
- Qual a probabilidade de, numa amostra, ter 1 DS?
- Qual a probabilidade de, numa amostra, ter 1 DS e Mulher (F)?
- Qual a probabilidade de, numa amostra, ter 1 DS dado que seja mulher (F)?

Condicionalidade

- Na primeira e segunda pergunta temos uma **probabilidade simples**, $P(M)$ e $P(DS)$.

- Na terceira pergunta temos uma **probabilidade de eventos combinados**, $P(M \cap DS)$.

- Na quarta pergunta temos uma **probabilidade condicional**, $P(DS|M)$.

Condicionalidade

- A **probabilidade condicional** é um segundo evento de um conjunto de eventos que ocorre após um evento já tenha ocorrido.

- Por exemplo, o primeiro evento eu conheço “ser mulher”, porém qual a probabilidade desta “pessoa ser um DS” dado que o primeiro evento ocorreu?

Teorema de Bayes

- No exemplo, para responder a quarta pergunta temos que fazer $P(DS|M) = \frac{\#\ de\ F\ e\ DS}{\#\ total\ de\ F}$.
- Temos que calcular o número de mulheres que são DS e dividir pelo total de mulheres.
- Este procedimento é denominado de teorema de Bayes:

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

Perguntas respondidas

- Qual a probabilidade de, numa amostra, ter 1 Mulher (F)?

$$P(F) = \frac{\#\ de\ F}{\#\ total} = \frac{24}{60}$$

- Qual a probabilidade de, numa amostra, ter 1 DS?

$$P(DS) = \frac{\#\ de\ DS}{\#\ total} = \frac{37}{60}$$

- Qual a probabilidade de, numa amostra, ter 1 DS e Mulher (F)?

$$P(DS \cap F) = \frac{\#\ de\ DS\ e\ F}{\#\ total} = \frac{10}{60}$$

- Qual a probabilidade de, numa amostra, ter 1 DS dado que seja mulher (F)?

$$P(DS|F) = \frac{P(DS \cap F)}{P(F)} = \frac{10/60}{24/60}$$

Voltando ao nosso problema

- Qual a probabilidade de um novo cliente ser Opt-In?

- Qual a probabilidade de um novo cliente ser Opt-In dado que está como Opt-Out?

- Desafio implícito: quantas regras teremos que criar para organizar esses bancos e termos uma classificação única (Opt-In, Opt-Out e Drop-Out) para cada cliente?

Resolução 1

- Desafio implícito: quantas regras teremos que criar para organizar esses bancos e termos uma classificação única (Opt-In, Opt-Out e Drop-Out) para cada cliente?

Gerenciamento de entrada 1	Gerenciamento de entrada 2	Gerenciamento de saída 1	Gerenciamento de saída 2
Opt-In	Opt-In	Unsubscribe	Unsubscribe
Opt-Out	Opt-Out	-	-

- Neste exemplo temos um problema de combinação, uma vez que a ordem do gerenciamento de entrada não importa. Além disto, não há como um cliente dar unsubscribe de um e-mail não recebido.
- Temos n = 8 elementos a serem arranjados em k = 4 posições.
- Deste modo, teríamos que criar $C_4^8 = \frac{8!}{4!(8-4)!} = 70$ regras para classificar todos os 40k clientes em Opt-In, Opt-Out e Drop-Out.

Resolução 2 e 3

- Qual a probabilidade de um novo cliente ser Opt-In?
- Qual a probabilidade de um novo cliente ser Opt-In dado que está como Opt-Out?

- Após a criação das regras e classificação dos clientes para os anos de 2018 e 2019 temos a seguinte tabela:

	Opt-In (2019)	**Opt-Out (2019)**	**Total**
Opt-In (2018)	7.000	5.000	12.000
Opt-Out (2018)	3.000	25.000	28.000
Total	10.000	30.000	**40.000**

Resolução 2

- Qual a probabilidade de um novo cliente ser Opt-In?

- Sabemos que o total de clientes são 40k e temos 10k como Opt-In, então:

$$P(In) = \frac{10k}{40k} = 0{,}25$$

Resolução 3

- Qual a probabilidade de um novo cliente ser Opt-In dado que está como Opt-Out?

- Aqui teremos que usar a regra da probabilidade condicional:

	Opt-In (2019)	Opt-Out (2019)	Total
Opt-In (2018)	7.000	5.000	12.000
Opt-Out (2018)	3.000	25.000	28.000
Total	10.000	30.000	**40.000**

Cuidado, este grupo aqui é Drop-Out

Clientes Opt-Out em 2018, mas desejaram ser Opt-In em 2019

Clientes Opt-Out em 2018 e continuaram Opt-Out em 2019

- $P(In|Out) = \frac{P(In \cap Out)}{P(Out)} = \frac{3k/40k}{25k/40k} = 0{,}12$

Exercícios

- Assistir o vídeo abaixo:
- https://www.youtube.com/watch?v=_sY3ZRxhBaM
- Quais aplicações que foram apresentadas em que probabilidades puderam ser aplicadas?

Exercícios

- Dada a seguinte tabela responda:

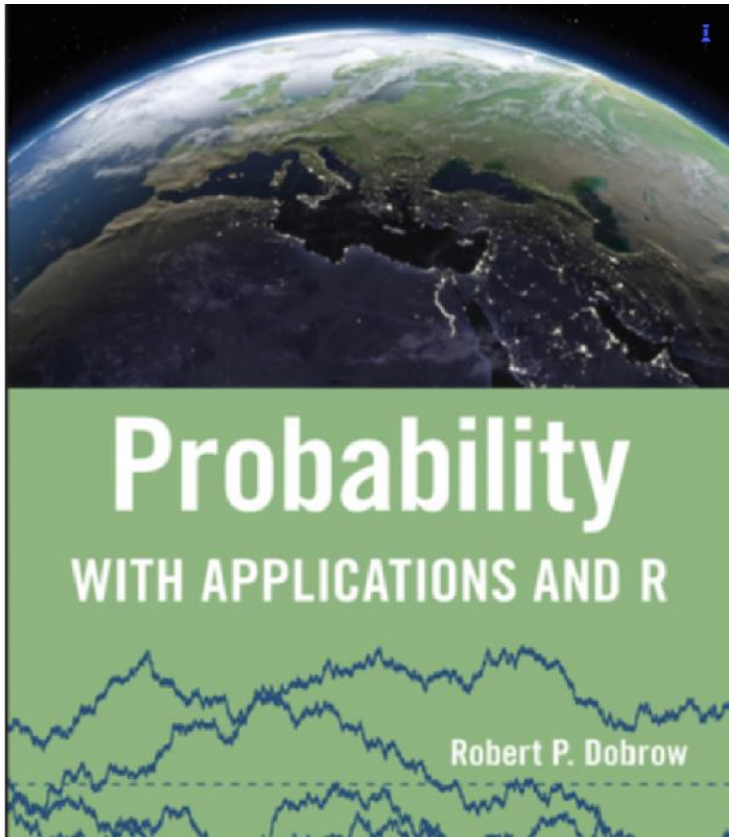
	1º ano	2º ano	3º ano
Homens	30	32	29
Mulheres	42	37	33

- Dado um sorteio:
- a) Qual a probabilidade de um aluno ser homem?
- b) Qual a probabilidade de um aluno ser do 2º ano?
- c) Qual a probabilidade de um aluno ser mulher dado que é do 3º ano?

- Parte das respostas:
- https://www.youtube.com/watch?v=EXJuqamzj4Y

Onde estudar mais!!

- Leitura



- Vídeos

- Intro. Prob.: https://pt.khanacademy.org/math/statistics-probability/probability-library/basic-theoretical-probability/v/basic-probability?modal=1

- Permutação: https://pt.khanacademy.org/math/statistics-probability/counting-permutations-and-combinations/permutation-lib/v/permutation-formula?modal=1

- Combinação: https://pt.khanacademy.org/math/statistics-probability/counting-permutations-and-combinations/combinations-lib/v/combination-formula?modal=1

- Teorema de Bayes: https://pt.khanacademy.org/math/ap-statistics/probability-ap/stats-conditional-probability/v/bayes-theorem-visualized

THE WAY WE ARE