Amostragem e Discretização

PUCRS
Pontifícia Universidade Católica do Rio Grande do Sul

– Aula 15 –
Coleta, Preparação e Análise de Dados

Prof. Me. Lucas R. C. Pessutto

Slides adaptados do material do Prof. Lucas Silveira Kupssinskü e do Prof. Luan Fonseca Garcia

Amostragem

- O processo de amostragem consiste em selecionar apenas um subconjunto dos dados para fazer alguma análise

- A motivação para amostrar dados em data mining é diferente da estatística
  - As vezes re-amostramos os dados apenas devido aos requerimentos de processamento e memória.

Amostragem

- Queremos amostras *representativas*
  - Propriedades semelhantes às do conjunto de dados

- Como amostragem é um processo estatístico, precisamos escolher a forma de amostrar que melhor vai manter as características que tentamos analisar.

Tipos de Amostragem

- Amostragem Aleatória Simples
  - Com ou sem reposição

Alguém consegue pensar em um caso onde essa abordagem não é adequada?

Tipos de Amostragem

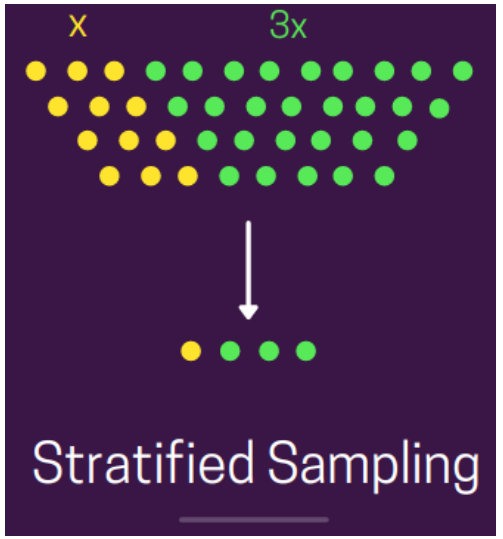
- Amostragem Aleatória Simples
  - Com ou sem reposição
- Amostragem Estratificada
  - Mantendo ou não a proporção dos estratos (normalmente definidos artificialmente)
- Amostragem por Conglomerado
  - Mantendo ou não a proporção dos conglomerados (grupos)

Amostragem Estratificada

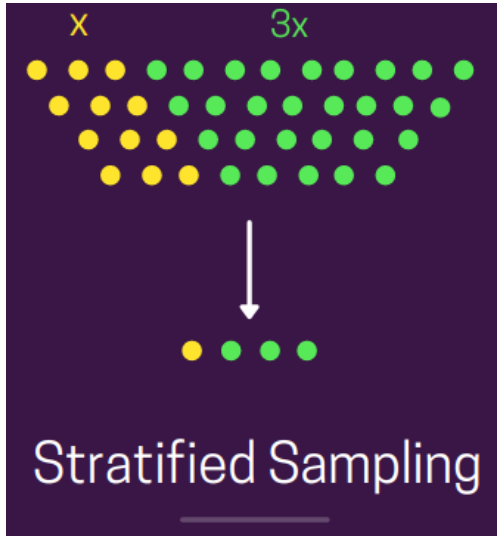
Strata (Age)	Number of People in Population	Number To Be Included in Sample
20-24	30,000	150
25-29	70,000	350
30-34	40,000	200
35-39	30,000	150
40-44	20,000	100
>44	10,000	50
Total	200,000	1,000

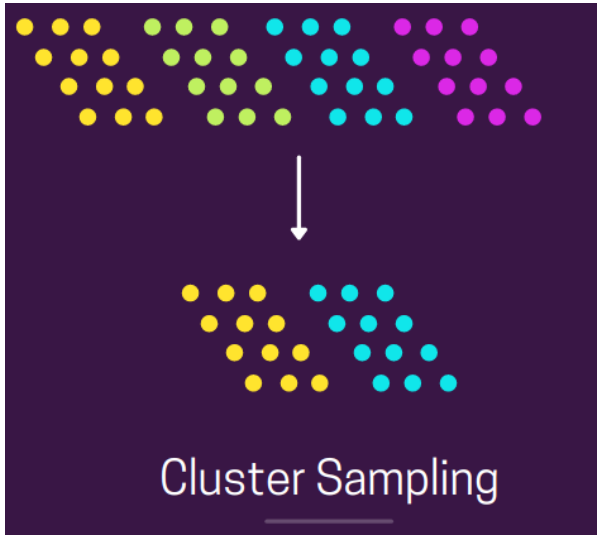
Amostragem Estratificada

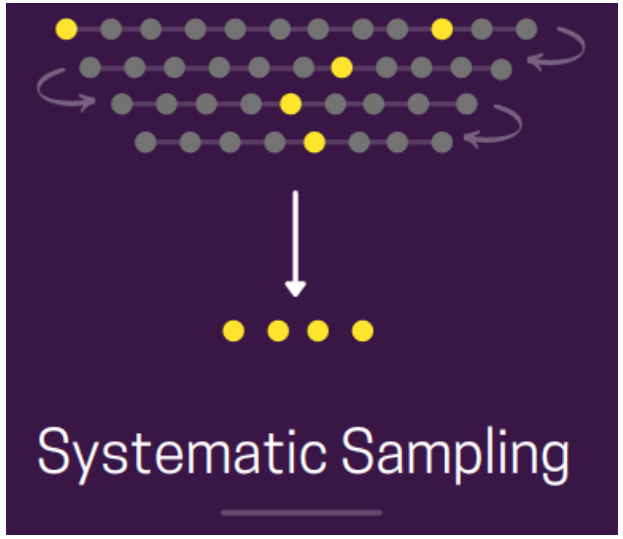
- Amostragem Estratificada Proporcional
  - Usada quando queremos manter a proporção dentro dos estratos.
- Amostragem Estratificada Desproporcional
  - Utilizada em conjuntos de dados desbalanceados.

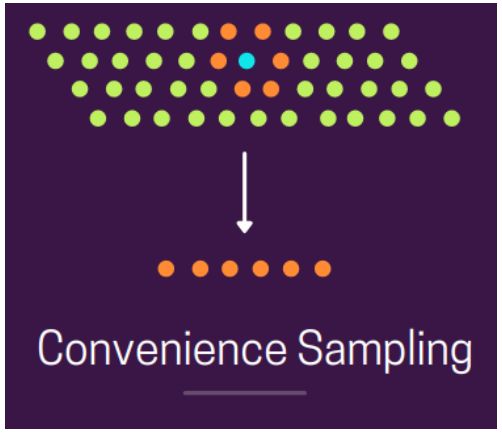


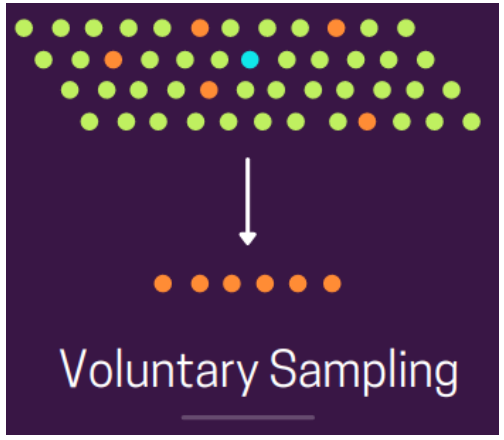
Resumo

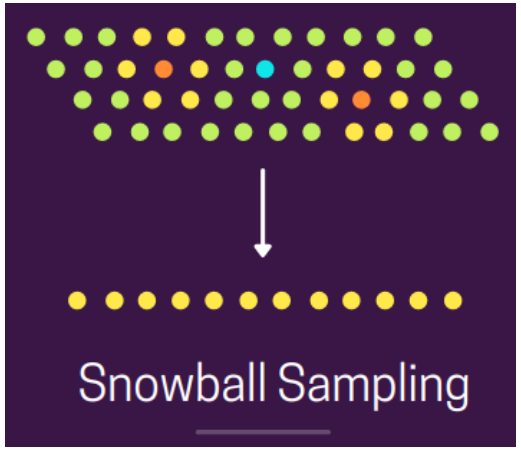












Fonte das ilustrações: https://towardsdatascience.com/8-types-of-sampling-techniques-b21adcdd2124

Binarização e One-hot Encoding

Valor Categórico	Valor Inteiro	Valores Binarizados			One-Hot Encoding				
		x_1	x_2	x_3	x_1	x_2	x_3	x_4	x_5
Péssimo	0	0	0	0	0	0	0	0	1
Ruim	1	0	0	1	0	0	0	1	0
Regular	2	0	1	0	0	0	1	0	0
Bom	3	0	1	1	0	1	0	0	0
Ótimo	4	1	0	0	1	0	0	0	0

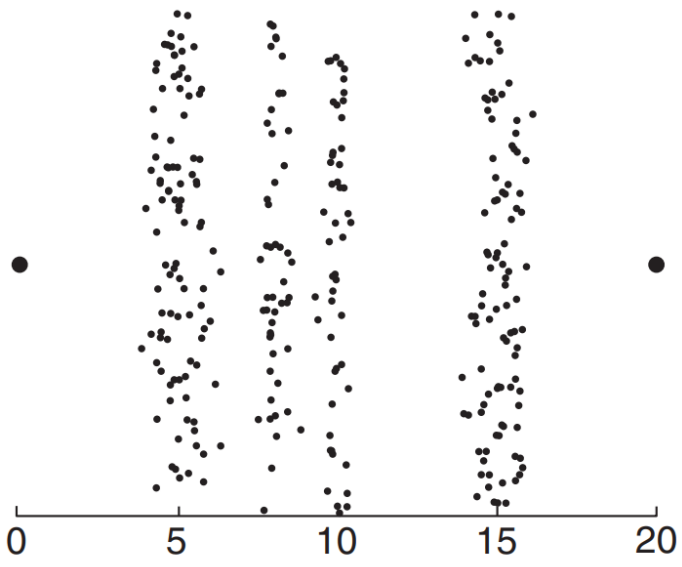
Observação: Eventualmente pode ser necessário codificar um único atributo binário como duas *features* assimétricas. Por exemplo para codificar o sexo de uma pessoa. A representação binária assimétrica é ineficiente em termos de custo de memória.

Discretização de atributos contínuos

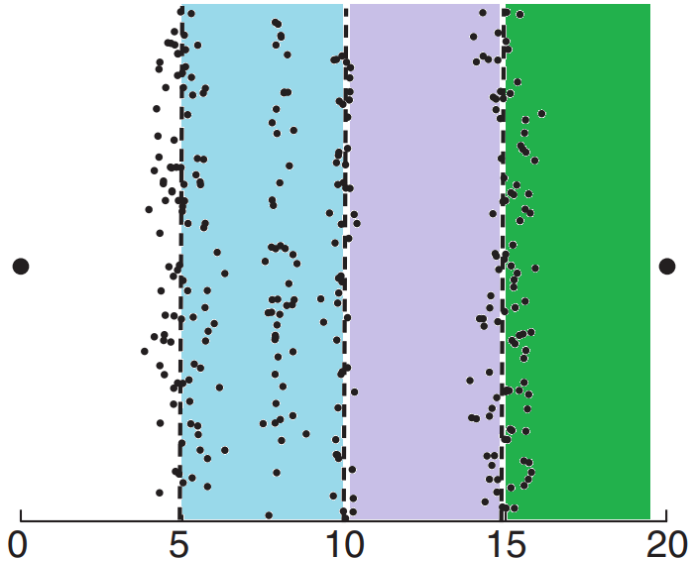
- Envolve dois passos:
  - Decidir um número de categorias para dividir o atributo;
  - Definir uma função de mapeamento do valor contínuo para uma categoria.

- Podemos representar a discretização usando uma série de intervalos: $\{(k_0, k_1], (k_1, k_2], \dots, (k_{n-1}, k_n)\}$

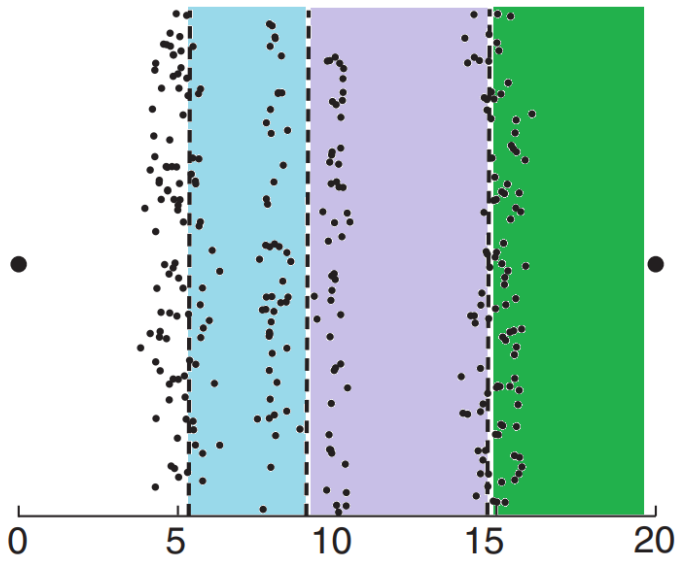
Discretização de atributos contínuos



(a) Original data.



(b) Equal width discretization.



(c) Equal frequency discretization.

Atributos categóricos com muitas variações

- Caso um atributo categórico possua muitos valores pode ser necessário diminuir o número de categoriais;
  - Para variáveis ordinais podemos aplicar “discretização”;
  - Para variáveis nominais temos que aplicar algum conhecimento de domínio.

Exercício Avaliativo

- (Amostragem) Escreva o corpo da função mostrada abaixo:

```
amostragem_aleatória(dataframe, amostras, reposicao=True)
```

Essa função recebe três parâmetros:

- dataframe = pandas contendo os dados
- amostras = número de amostras desejadas
- reposicao = amostragem com ou sem reposição

Essa função deve retornar um dataframe com o número exato de amostras

Exercício Avaliativo

- (Amostragem) Escreva uma função com a assinatura mostrada abaixo:

```
amostragem_estratificada(dataframe, amostras, coluna)
```

Essa função recebe três parâmetros:

- dataframe = pandas contendo os dados
- amostras = número de amostras desejadas
- coluna = nome de uma coluna do dataset

Essa função deve retornar um dataframe com o número exato de amostras, mantendo a proporção de valores da coluna escolhida presente no dataframe original.

Exercício Avaliativo

- Exemplo:

Dataset Entrada (100 registros)

Valor da Coluna	Qtde Valores
A	50
B	30
C	10
D	20

Dataset Saída (10 registros)

Valor da Coluna	Qtde Valores
A	5
B	3
C	1
D	2

Exercício Avaliativo

- Usando o dataset “alugueis” fornecido, realize testes nas funções desenvolvidas:
  - Crie um dataset com apenas 100 amostras
  - Crie outro dataset com 200 amostras, preservando a proporção de imóveis em cada cidade
- Faça a discretização do atributo área do imóvel em três categorias:
  - “PEQUENO” = até $50m^2$
  - “MÉDIO” = 50 até $100m^2$
  - “GRANDE” = > $100m^2$
- Divida o atributo total em 5 categorias e utilize one-hot encoding para discretizá-lo