Pré-Processamento de Dados

ACH5504 – Mineração de Dados

Notas de aulas baseadas no livro

"*Introduction to Data Mining*"

Tan, Steinbach, Karpatne, Kumar

Resumo

- Agregação
- Amostragem
- Redução de Dimensionalidade
- Seleção de subconjunto de atributos
- Criação de atributos
- Discretização e Binarização
- Transformação de Atributo

Agregação

- Combinar dois ou mais atributos (ou objetos) em um único atributo (ou objeto)

- Propósito
  - Redução de dados
    - Redução de número de atributos ou objetos
  - Mudança de escala
    - Cidades agregadas em regiões, estados, países, etc.
    - Dias agregados em semanas, semanas, meses ou anos
  - Dados mais “estáveis”
    - Dados agregados tendem a ter menos variabilidade

Exemplo: precipitação na Austrália

- Este exemplo é baseado na precipitação em Austrália do período 1982 a 1993.

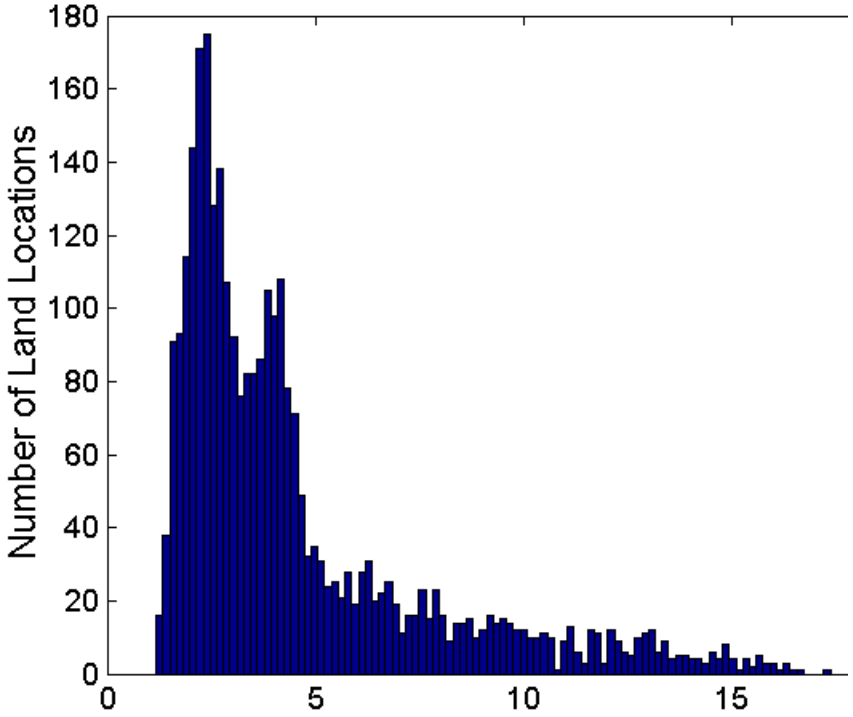
  O próximo slide mostra

  - Um histograma para o desvio padrão da precipitação mensal média para 3030 células de grade (0,5° × 0,5° cada) na Austrália, e
  - Um histograma para o desvio padrão da precipitação anual média para os mesmos locais.

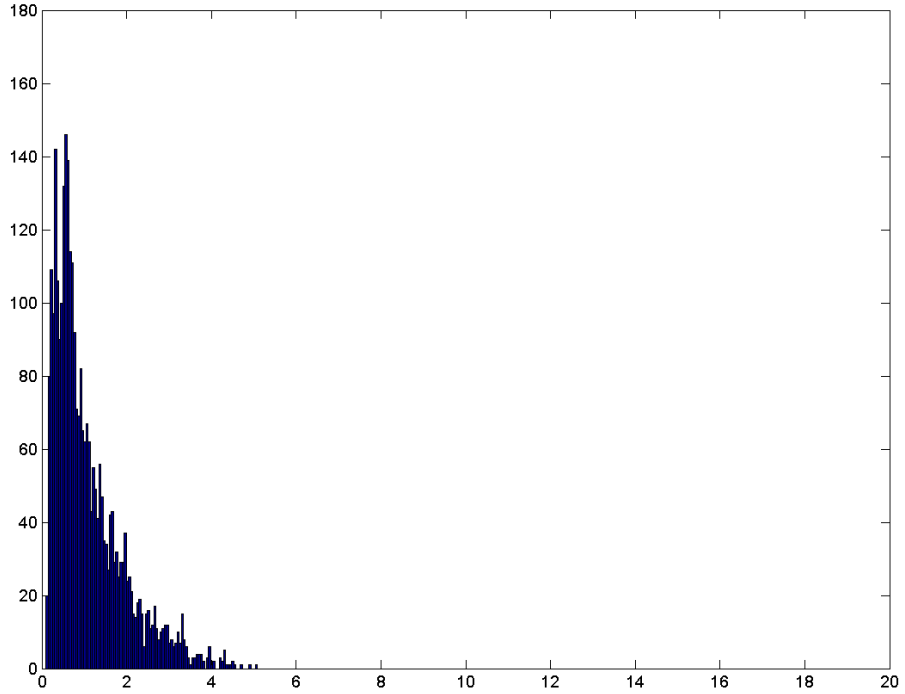
- A precipitação anual média tem menos variabilidade do que a precipitação mensal média.
- Todas as medições de precipitação (e seus desvios padrão) são em centímetros.

Exemplo: precipitação na Austrália …

Variação da precipitação em Austrália



Desvio padrão da precipitação mensal média



Desvio padrão da precipitação anual média

Amostragem

- A amostragem é a técnica principal empregada para a redução de dados.
  - É usado frequentemente para a investigação preliminar e a análise final dos dados.

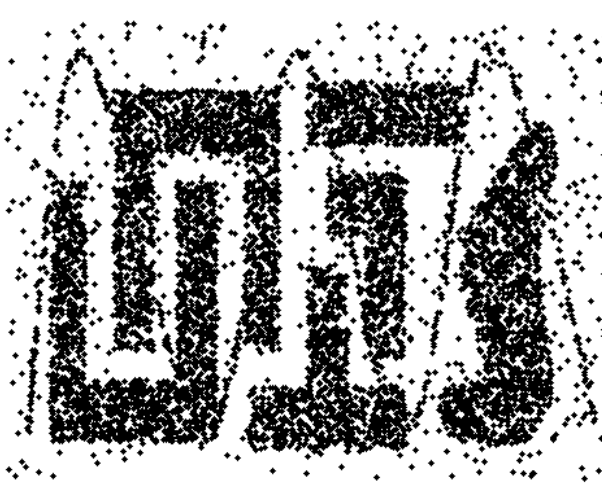
- Estatísticos muitas vezes aplicam a amostragem porque a obtenção de conjunto completo de dados é caro e demorado.

- A amostragem é tipicamente usada na mineração de dados porque o processamento de conjunto de dados completos é caro e demorado.

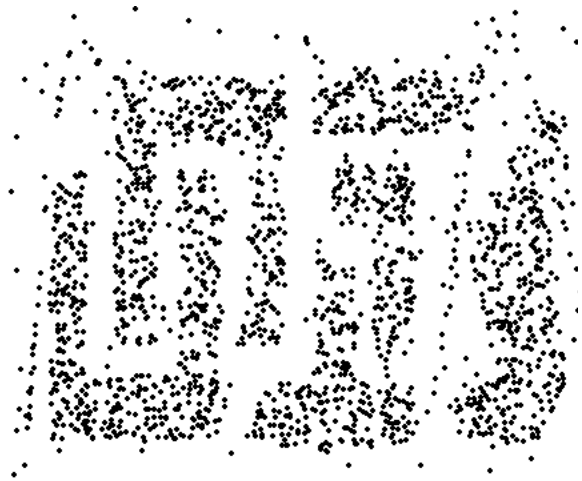
Amostragem …

- A chave principal para uma amostragem eficaz é o seguinte:
  - Uma amostra funcionará quase tão bem quanto todo o conjunto de dados, se a amostra for representativa
  - Uma amostra é representativa se ela tem aproximadamente as mesmas propriedades (de interesse) como o conjunto original de dados

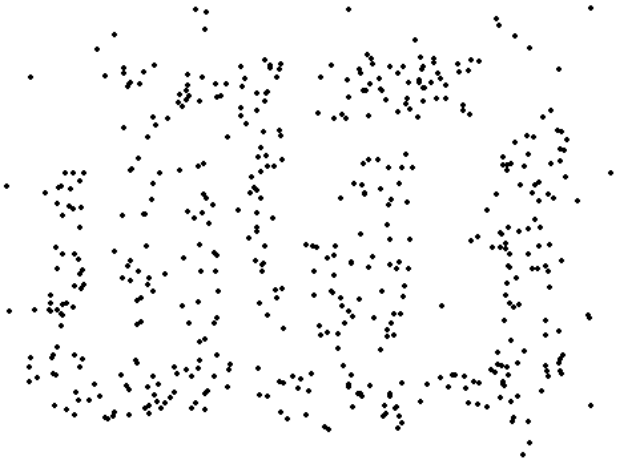
Tamanho da amostra



8000 pontos



2000 Pontos



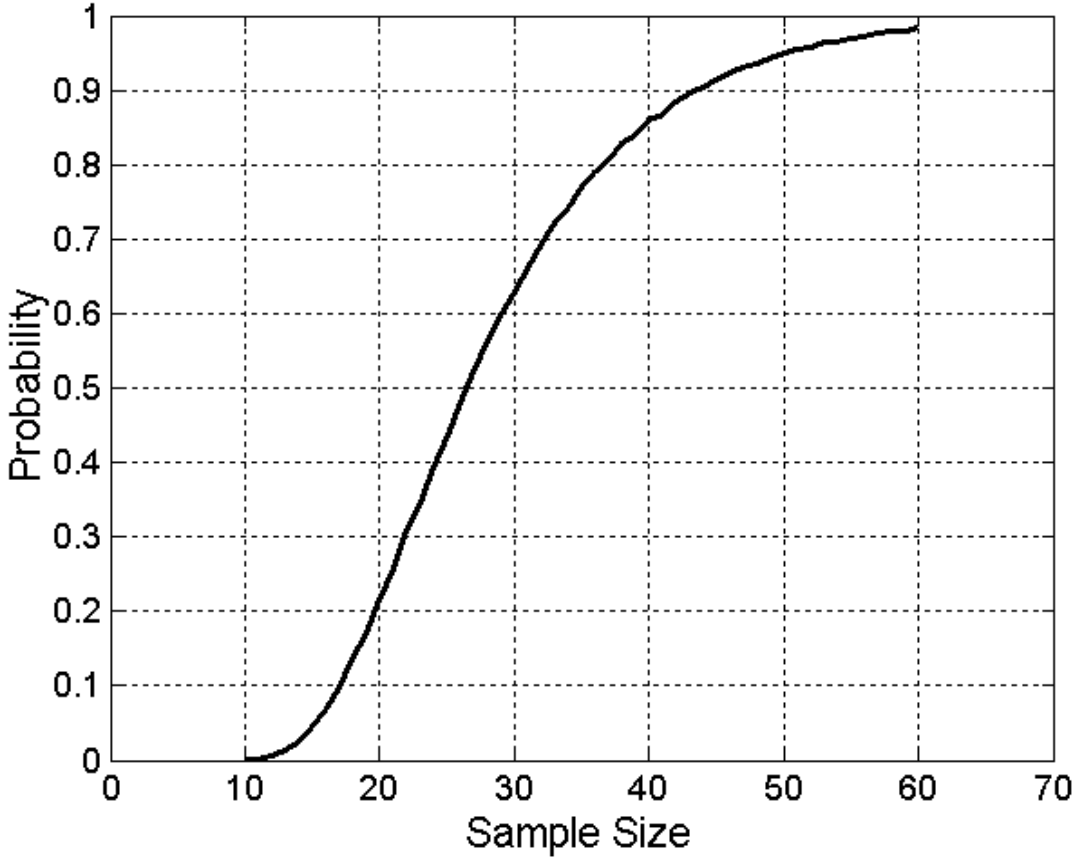
500 Pontos

Tipos de amostragem

- Amostragem aleatória simples
  - Cada instância tem mesma probabilidade
  - Sem reposição
    - Cada instância selecionada é retirada da população
  - Com reposição
    - Instâncias não são removidas da população.
    - A mesma instância pode ser escolhida mais do que uma vez
- Amostragem estratificada
  - Dados são particionados e instâncias são sorteadas de cada partição

Tamanho da amostra

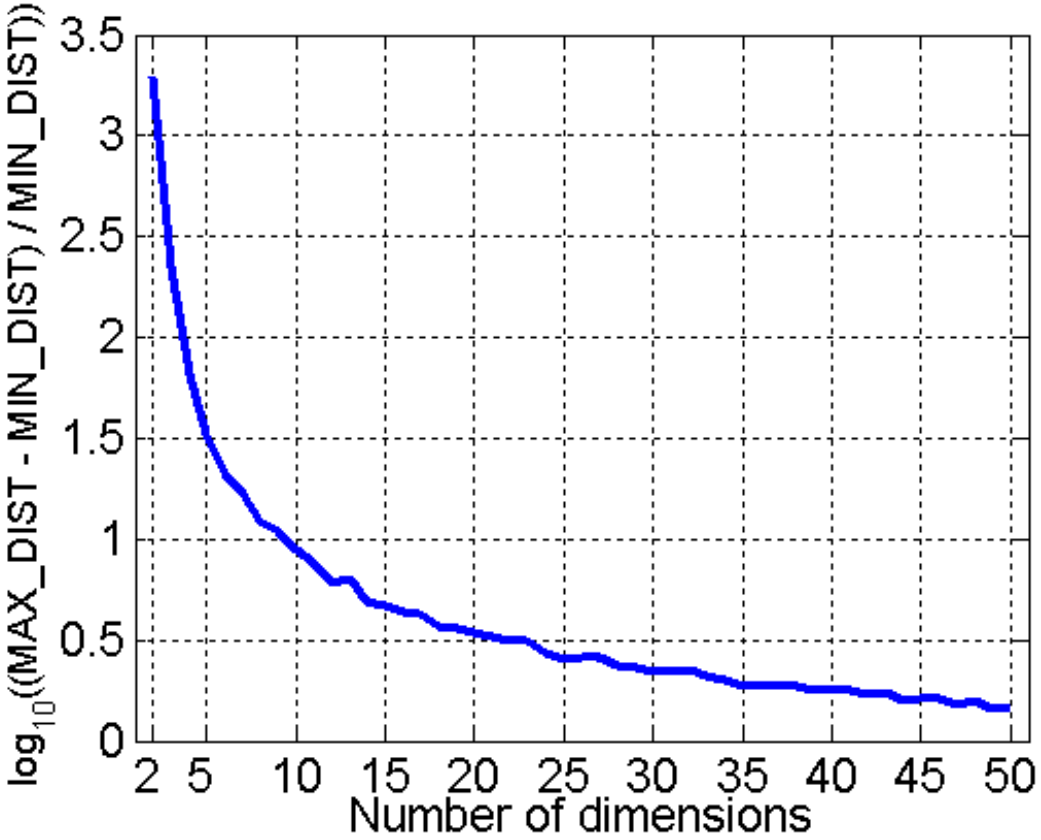
- **Qual tamanho da amostra é necessário para obter pelo menos uma instância de dados para cada de 10 grupos de tamanho igual?**



Maldição da dimensionalidade

- Aumento do número de dimensionalidades deixa dados mais esparsos no espaço que ocupa.

- Definições de densidade e distância entre pontos, que são essenciais para a deteção de clusters e outlier, tornam-se menos significativas



- **Gere aleatoriamente 500 pontos**
- **Calcule a diferença entre a distancia mínima e máxima entre cada par de pontos**

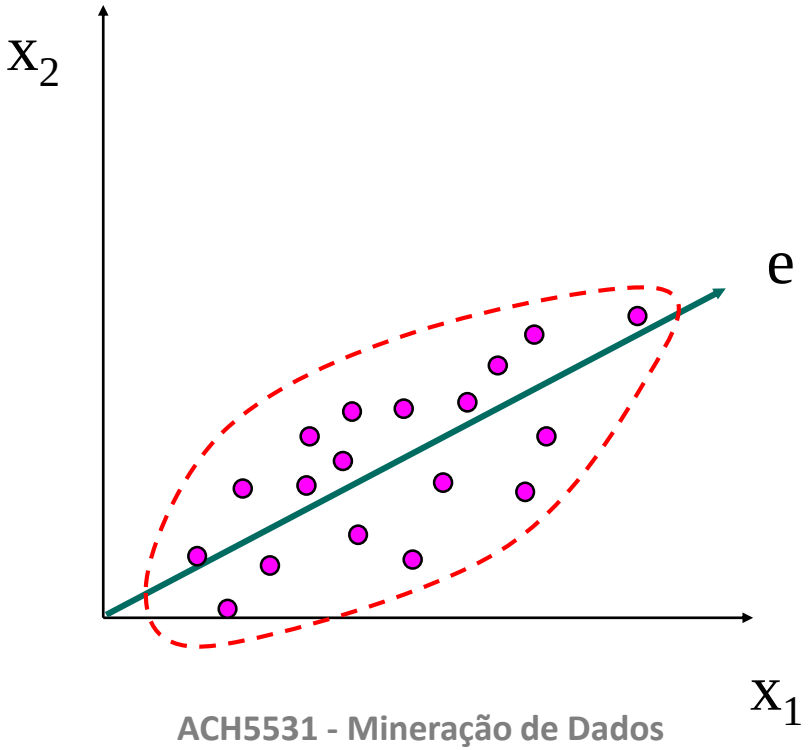
Redução de dimensionalidade

- Propósito:
  - Evitar a maldição da dimensionalidade
  - Reduzir a quantidade de tempo e a memória exigida pelos algoritmos de mineração de dados
  - Permitir que os dados sejam mais facilmente visualizados
  - Pode ajudar a eliminar atributos irrelevantes ou reduzir ruído

- Técnicas
  - Análise de componentes principais (Principal Components Analysis)
  - Decomposição de valor singular (Singular Value Decomposition)
  - Outras: técnicas supervisionadas e não-lineares

Redução de dimensionalidade: PCA

- O objetivo é achar uma projeção que captura a maior parte da variação dos dados.
- Dados originais são projetados em um espaço menos, composto pelos autovetores da matriz de co-variância



Redução de dimensionalidade: PCA

Seleção de subconjunto de atributos

- Outra maneira de reduzir a dimensionalidade dos dados
- Atributos redundantes
  - Duplica a maior parte ou todas as informações contidas em um ou mais atributos diferentes
  - Exemplo: preço de compra de um produto e o valor do imposto sobre vendas pago
- Atributos irrelevantes
  - Não contêm informações úteis para a tarefa de mineração de dados à mão
  - Exemplo: a identidade dos alunos é irrelevante para a tarefa de prever a nota média deles
- Muitas técnicas desenvolvidas, especialmente para a classificação

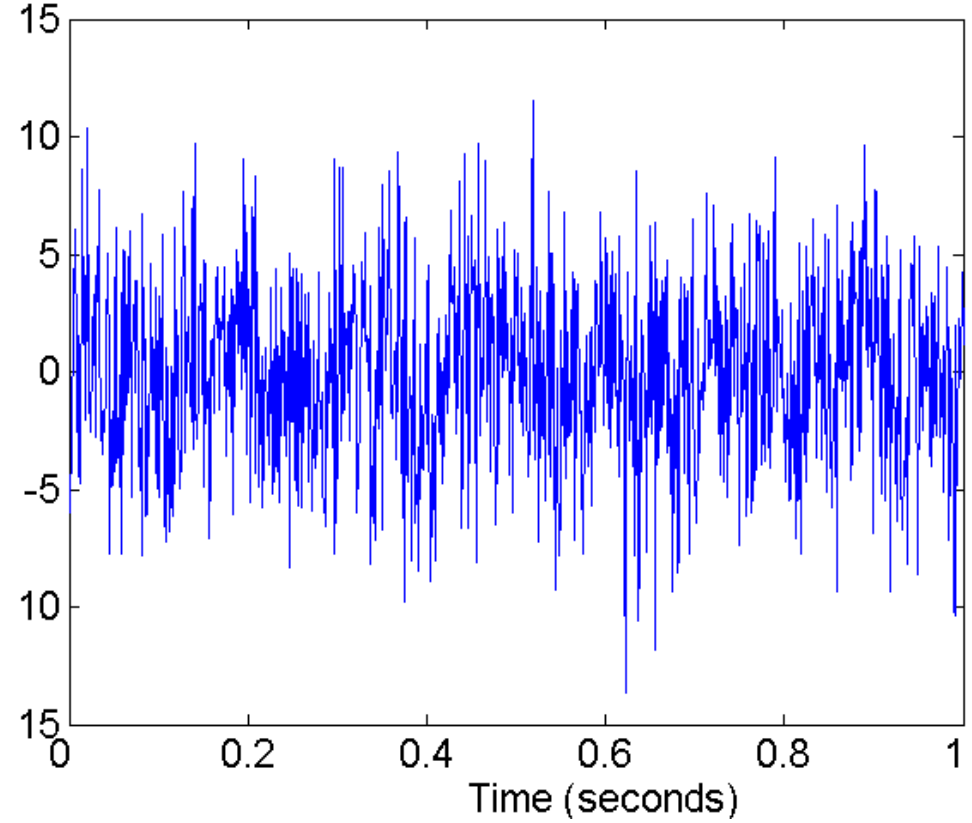
Criação de atributos

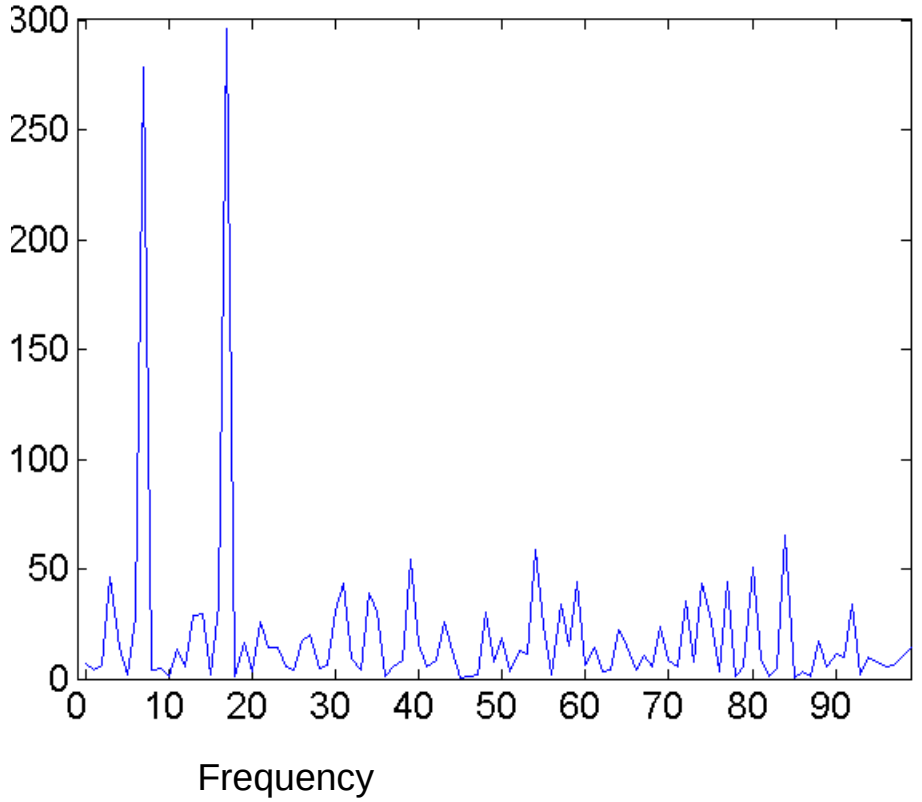
- Criar novos atributos que capturam a informação importante nos dados de forma mais eficiente que os originais

- Três metodologias gerais:
  - Extração de atributos (Feature extraction)
    - Exemplo: extração de bordas em imagens
  - Construção de atributos (Feature construction)
    - Exemplo: dividindo a massa por volume para obter a densidade
  - Mapeamento de dados em outo espaço
    - Exemplo: transformada de Fourier e analise de wavelet

Mapeamento de dados para um novo espaço

- **Transformada de Fourier e wavelet**





Duas ondas senoidais + ruído

Freqüência

Discretização

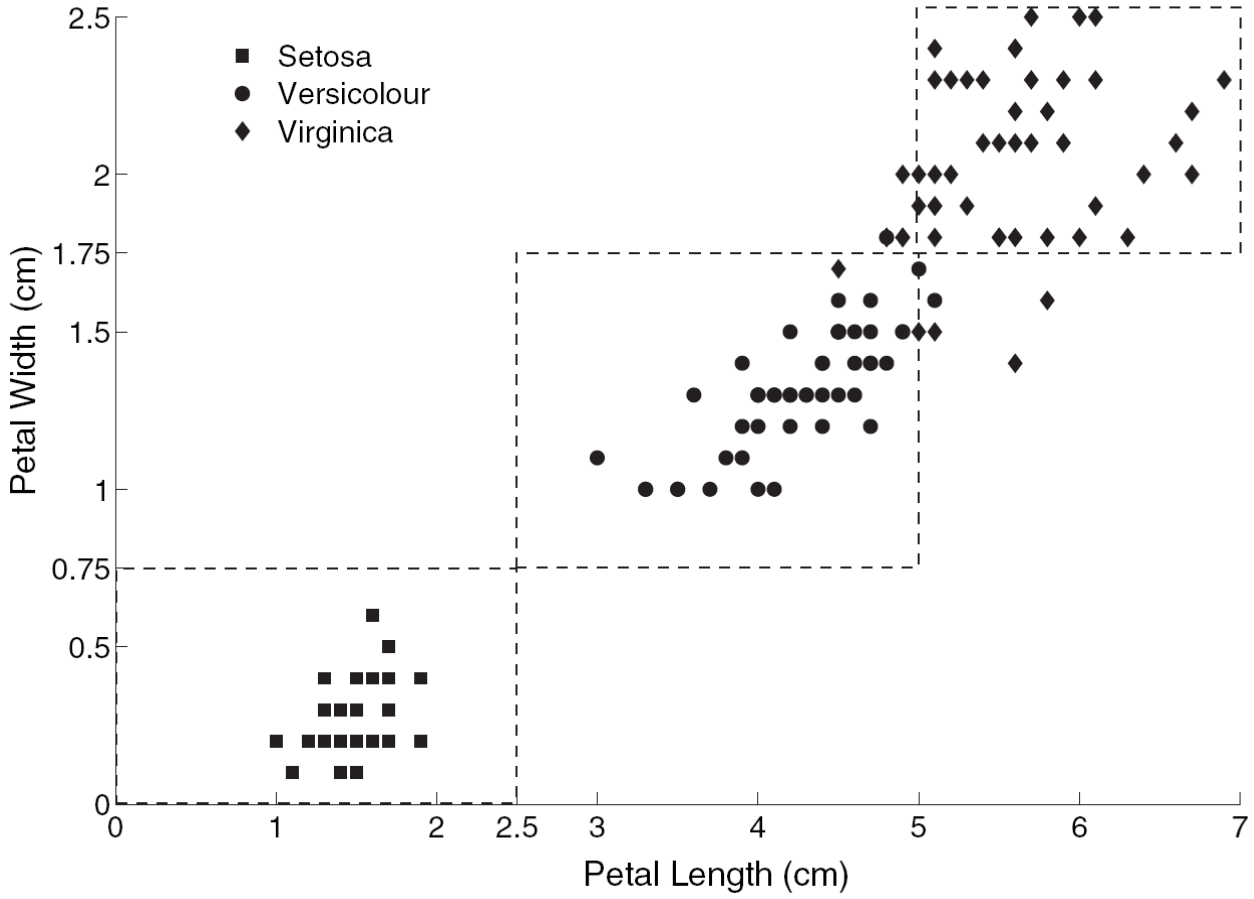
- Discretização é o processo de conversão de um atributo contínuo em um atributo ordinal
  - Um número potencialmente infinito de valores é mapeado para um pequeno número de categorias
  - A discretização é comumente usada na classificação
  - Muitos algoritmos de classificação funcionam melhor se ambas as variáveis independentes e dependentes tiverem apenas alguns valores
  - Damos uma ilustração da utilidade da discretização usando o conjunto de dados Iris

Conjunto de dados Iris

- Conjunto de dados da planta da íris.
  - Pode ser obtido a partir de site the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - Autor é Douglas Fisher
  - Três tipos da flor (classes):
    - Setosa
    - Versicolour
    - Virginica
  - Quatro (não-classe) atributos
    - Largura e comprimento de sépala
    - Largura e comprimento de pétala

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

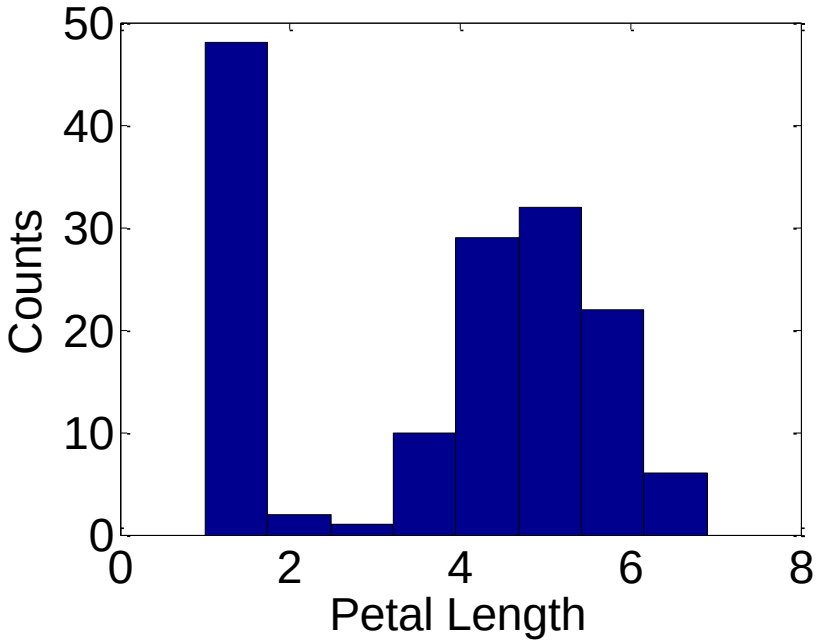
Discretização: Exemplo de Iris



Pétala de largura baixa ou comprimento baixo implica Setosa.
Pétala de largura média ou comprimento médio implica Versicolour.
Pétala de largura alta ou comprimento alto implica Virginica.

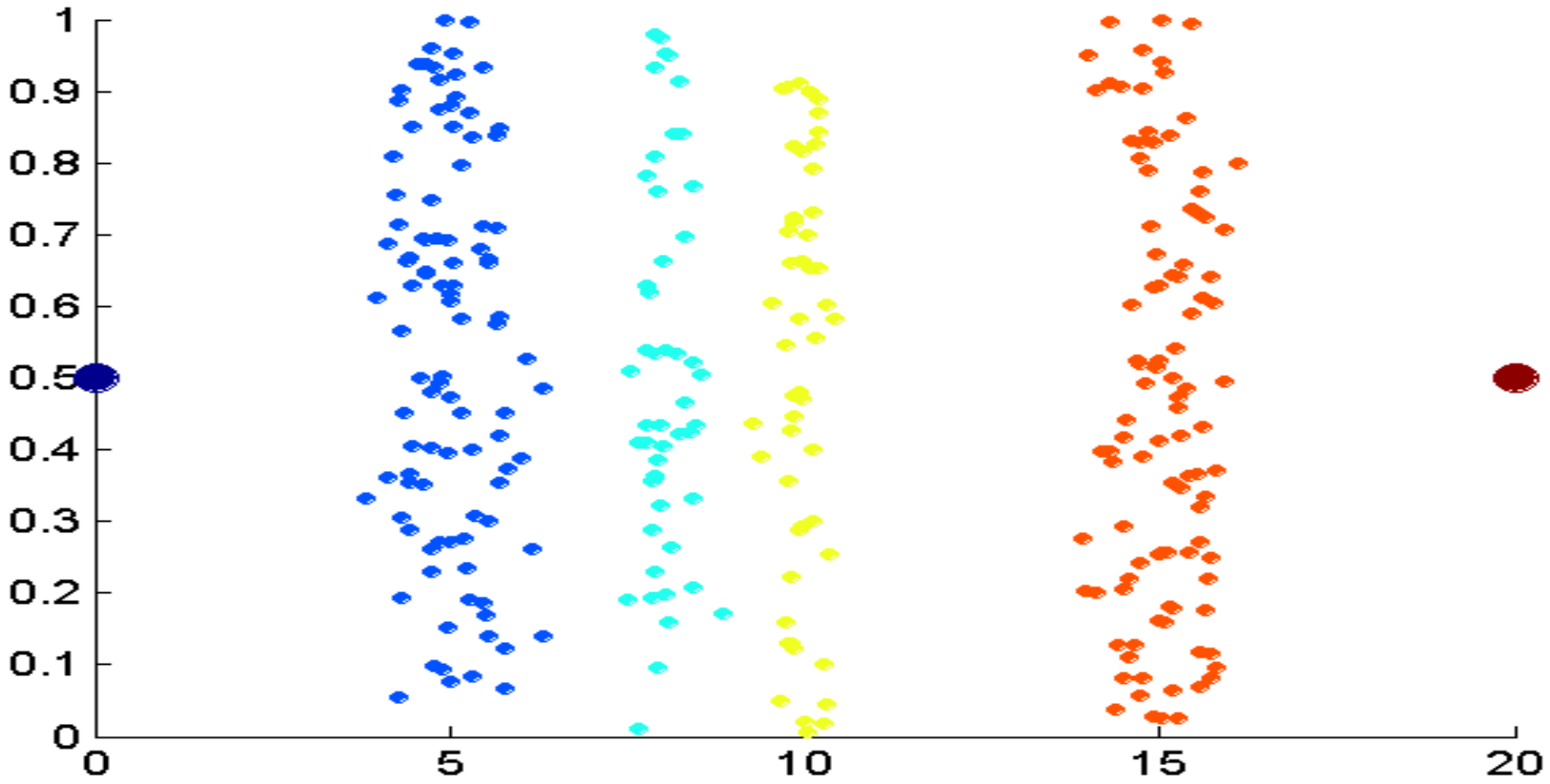
Discretização: Exemplo Iris …

- Como podemos dizer qual é a melhor discretização?
  - Discretização não supervisionada : encontrar quebras nos valores de dados
    - Exemplo: Comprimento de pétala



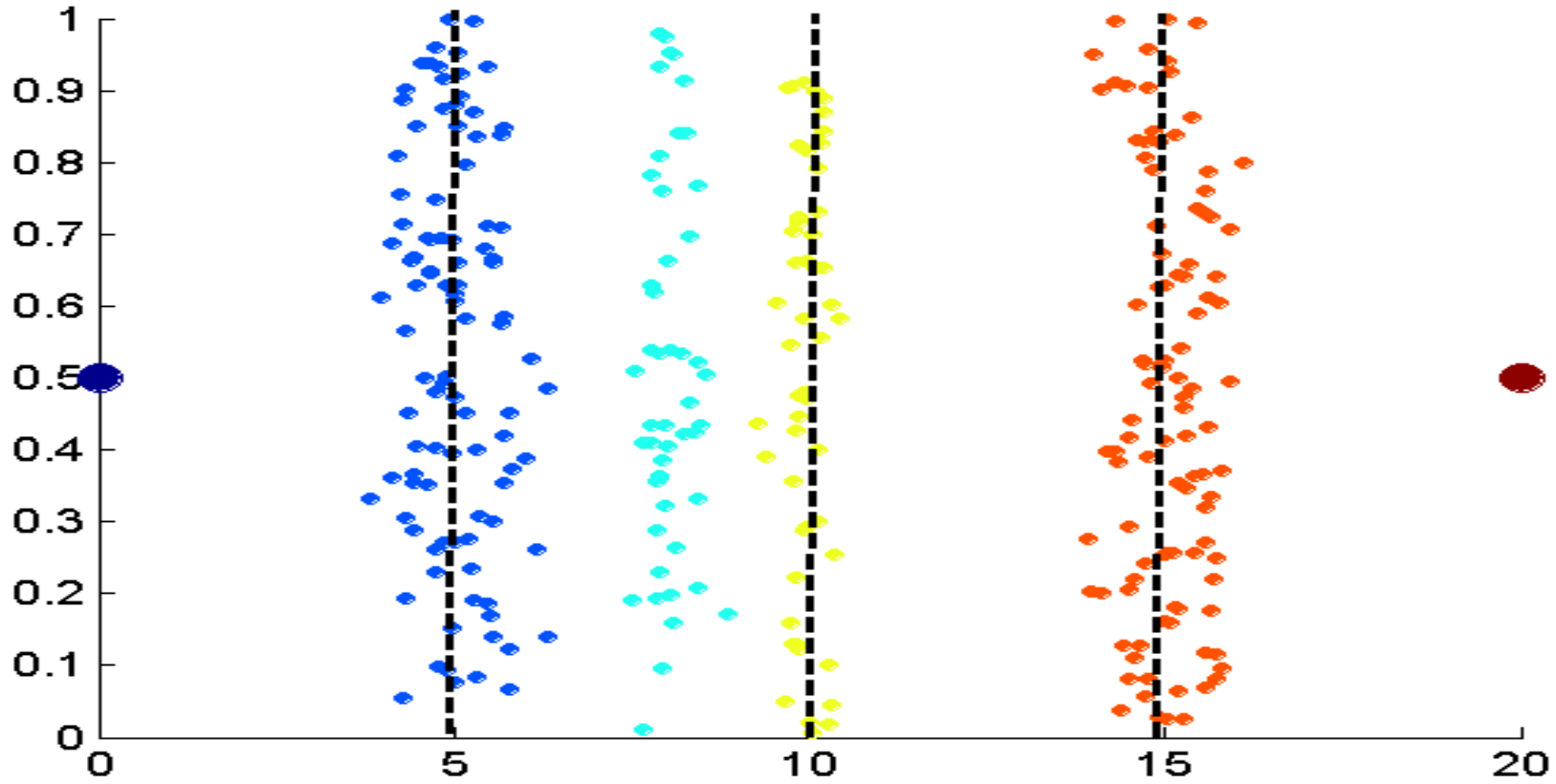
  - Disretização supervisionada: Usas etiquetas de classes para localizar quebras

Discretização sem usar etiquetas de classe



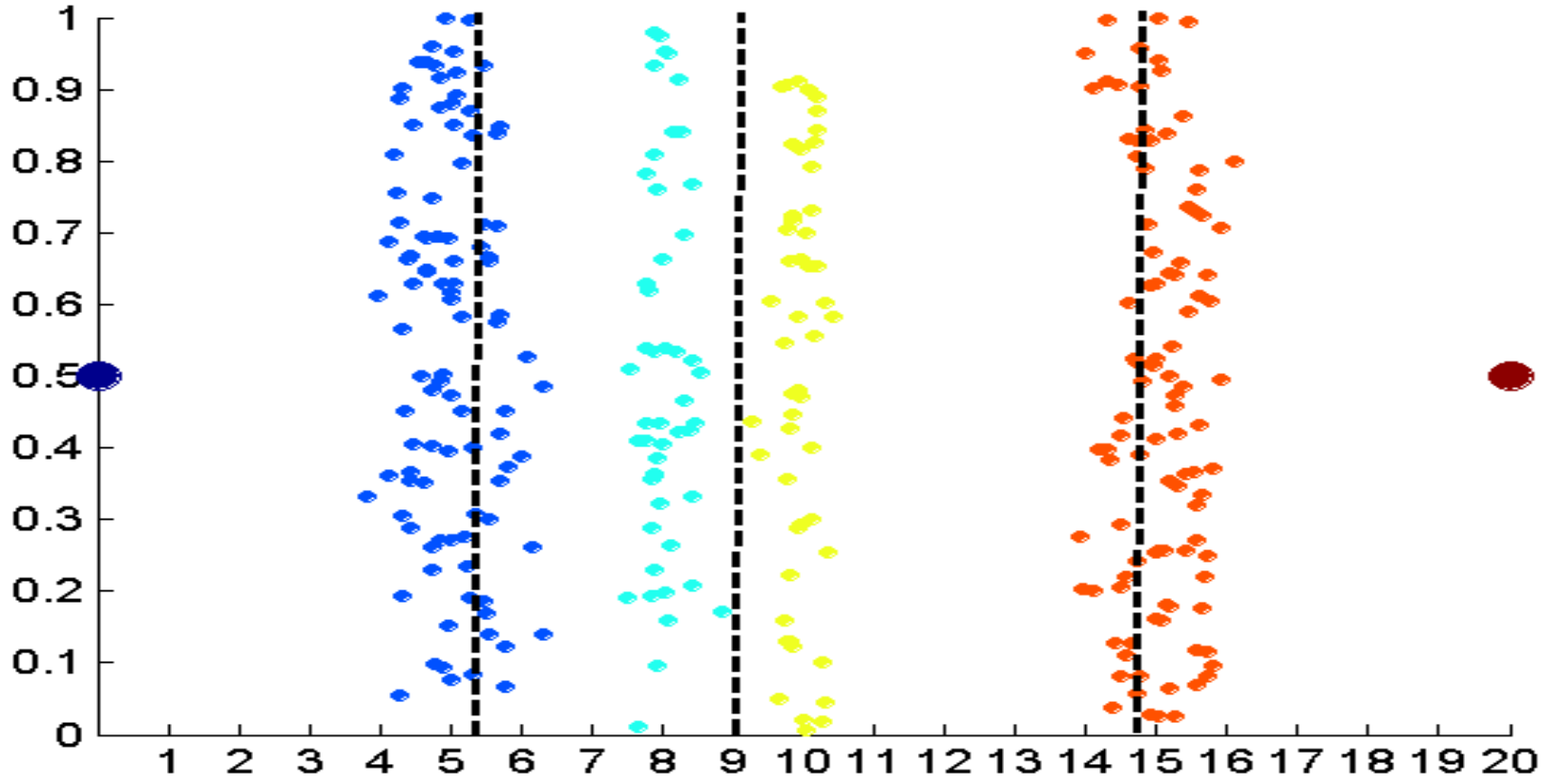
Os dados consistem quatro grupos de pontos e dois outliers. Os dados são unidimensionais, mas um componente aleatório y é adicionado para reduzir a sobreposição.

Discretização sem usar etiquetas de classe



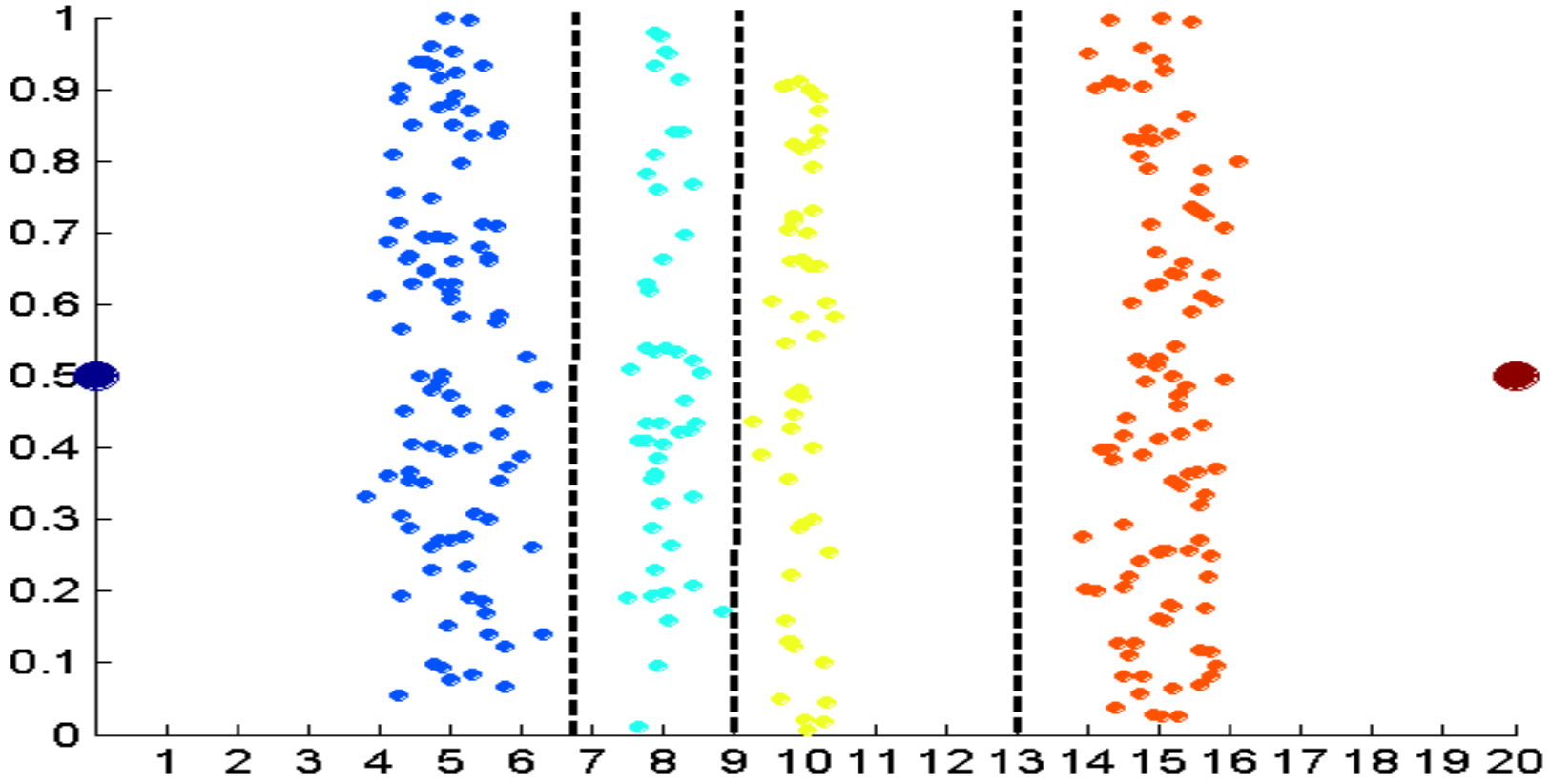
Abordagem de igual largura de intervalos foi usada para obter 4 valores.

Discretização sem usar etiquetas de classe



Abordagem de igual frequência foi usada para obter 4 valores.

Discretização sem usar etiquetas de classe



Abordagem de K-means foi usada para obter 4 valores.

Binarização

- Binarização mapeia um atributo contínuo ou categórico em uma ou mais variáveis binárias
- Tipicamente usado para análise de associação
- Frequentemente converte um atributo contínuo para um atributo categórico e, em seguida, converte um atributo categórico para um conjunto de atributos binários
  - Análise de associação precisa de atributos binários assimétricos
  - Exemplos: cor dos olhos e altura medidas como {baixa, média, alta}

Transformação de atributos

- Uma transformação de atributo é uma função que mapeia todo o conjunto de valores de um determinado atributo para um novo conjunto de valores de substituição, de forma que cada valor antigo pode ser identificado com um dos novos valores
  - Funções simples: x^k, log(x), e^x, |x|
  - Normalização ou padronização
    - Refere-se a várias técnicas para ajustar às diferenças entre os atributos em termos de frequência de ocorrência, média, variância, intervalo
    - Tira sinal indesejado, comum, por exemplo, sazonalidade
  - Na estatística, padronização refere-se subtração de média e divisão por desvio padrão

Tipos de Normalização

- **Normalização min-max**: para $[new\_min_A, new\_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Exemplo: Renda variando entre R$12.000 e R$98.000 para intervalo [0.0,1.0]. Assim o valor R$73.000 é mapeado para

    (1,0 – 0,0) (73.600 - 12.000) / (98.000 - 12.000) + 0,0 = 0,716

- **Normalização Z-score** (μ – média, σ – desvio padrão):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

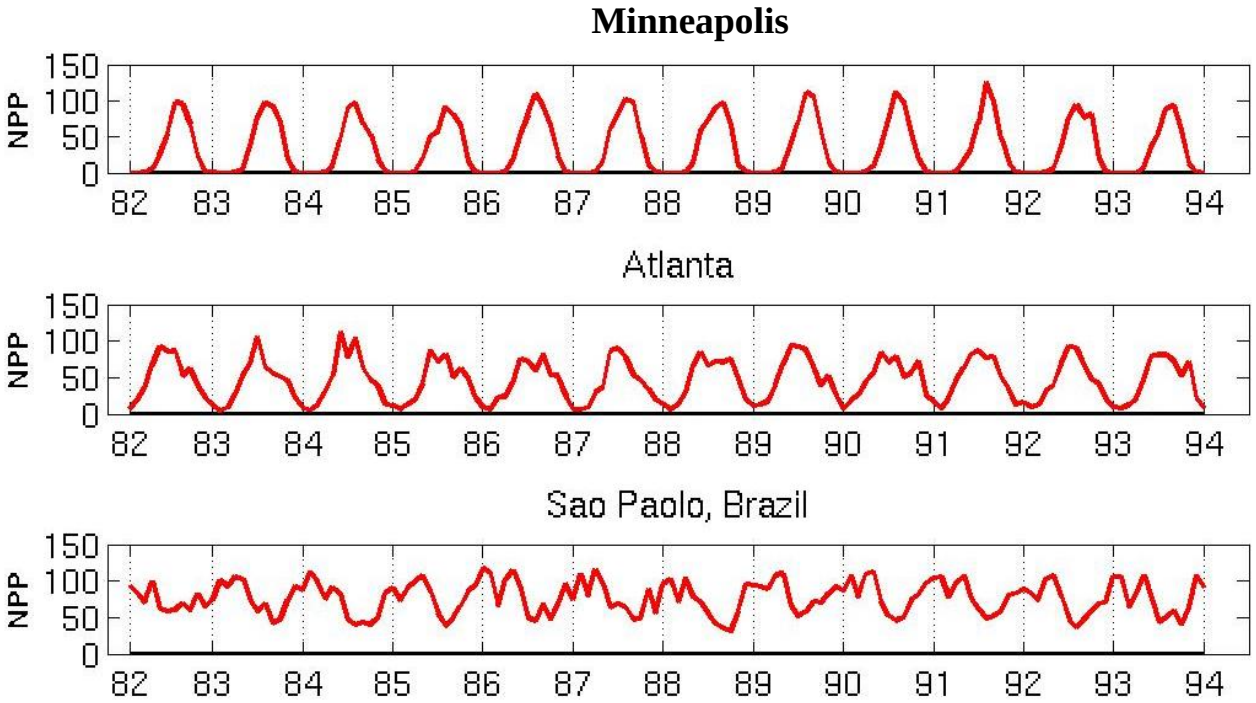
  - Exemplo: seja μ = 54.000, σ = 16.000, então (73.000-54.000)/16.000=1,225

- **Normalização por escalonamento de decimal**

$$v' = \frac{v}{10^j}$$

  - onde j é o menor inteiro tal que max(|v'|) < 1

Exemplo: Série temporal do crescimento de plantas

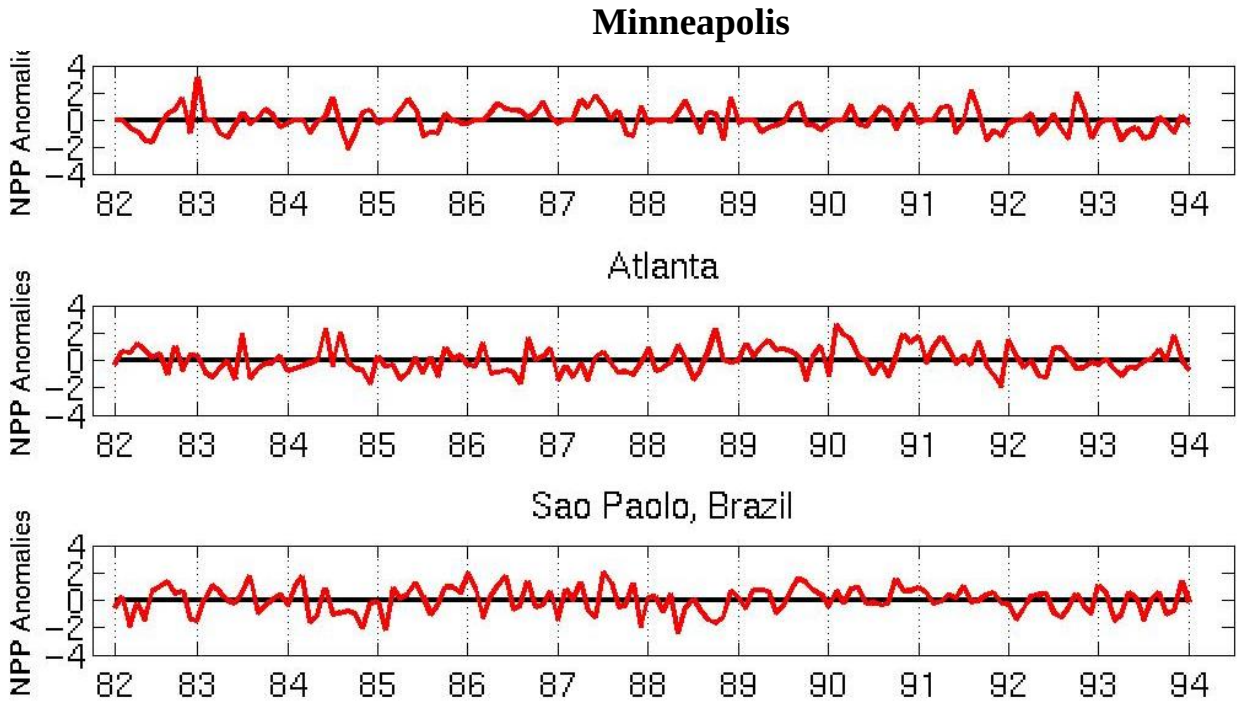


Net Primary Production (NPP) é uma medida de crescimento de plantas usada pelos cientistas do ecossistema.

Correlações entre séries temporais

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

Sazonalidade é responsável por maior correlação



Normalizado usando index **Z Score** mensal:

Subtraindo a média mensal e dividindo por desvio padrão mensal

Correlações entre séries temporais

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000