Entradas

Carlos Diego Rodrigues

5 de outubro de 2021

Universidade Federal do Ceará

Entrada: conceitos, instâncias, atributos

- Componentes de uma entrada para o aprendizado.
- O que é um conceito?
  - Regressão, Classificação, associação, agrupamento.
- O que é um exemplo?
  - Relações, arquivos, recursão
- O que é um atributo?
  - Nominal, ordinal, intervalo, valor
- Preparando a entrada

Componentes da entrada

- Conceitos: algo a ser aprendido. Nosso objetivo é construir um conceito operacional e inteligível, uma regra ou cálculo capaz de fornecer valores para as instâncias.
- Instâncias: Exemplos individuais de um conceito a ser aprendido.
- Atributos: medidas sobre uma instância.

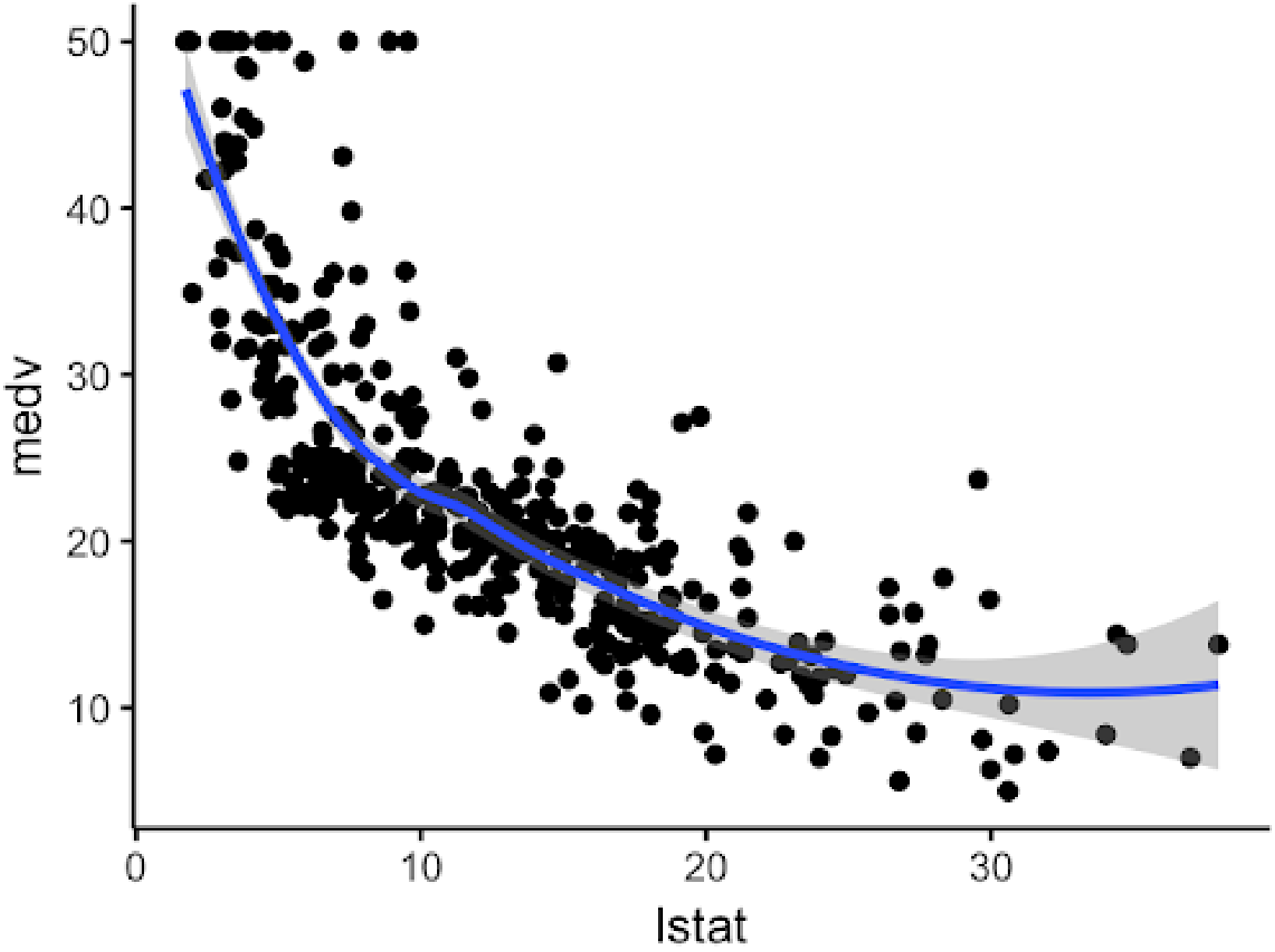
O que é um conceito

- Conceito: algo a ser aprendido.
- É a saída de um esquema de aprendizado.
- Tipos de aprendizado:
  - Regressão: predição de um valor numérico (contínuo).
  - Classificação: predição de uma classe discreta.
  - Associação: detecção de relações entre as características.
  - Agrupamento: reunião de instâncias em grupos por similaridade.

Regressão

- Tipo mais comum de mineração de dados.
- Busca-se uma função numérica capaz de descrever um conjunto de pontos em um espaço multidimensional.
- Regressão é aprendizado *supervisionado*.
- Medidas de distância (ex.: teste χ^2) calculam a taxa de sucesso de uma regressão.

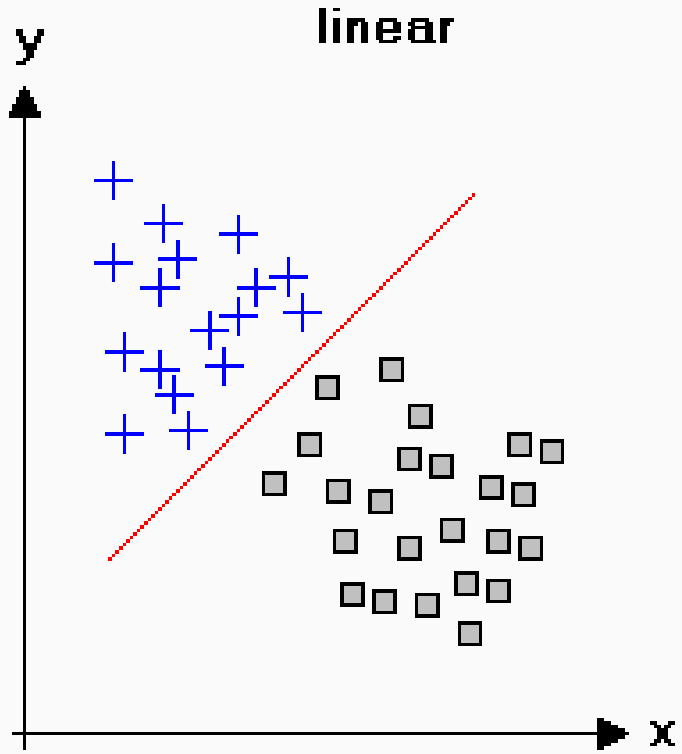
Exemplo regressão

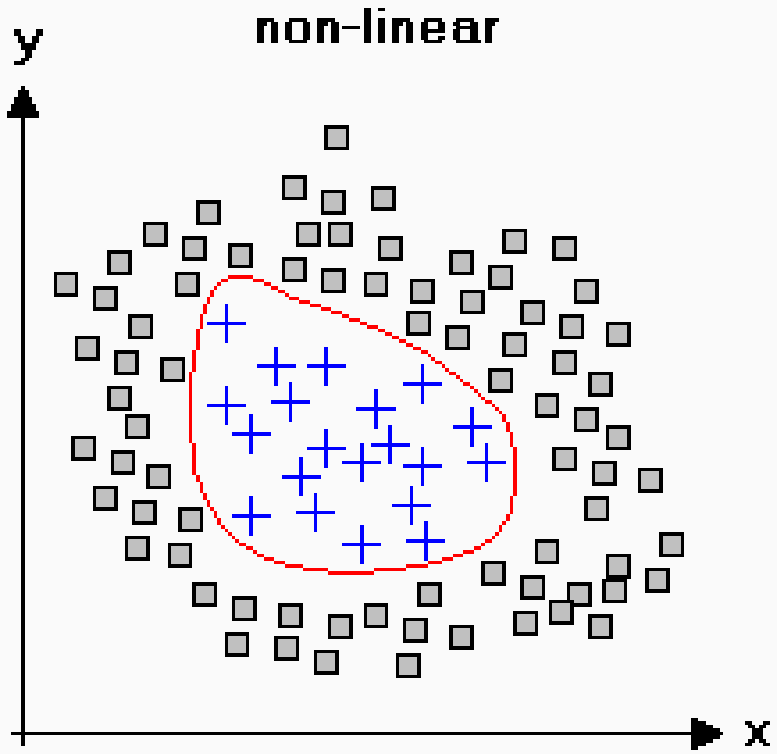


Classificação

- Exemplos: previsão do tempo, lentes de contato, classificação de espécies, negociações (empréstimos).
- Classificação é aprendizado *supervisionado*.
- Resultado é uma *classe* de um exemplo.
- A medida do sucesso é a proporção de instâncias corretamente classificadas em um conjunto de teste.
- Na prática a medida é subjetiva...

Exemplo classificação





Associação

- Pode ser aplicado quando não há classes ou não há estruturas consideradas interessantes.
- Associação é aprendizado *não supervisionado*.
- Diferença para classificação:
  - Pode predizer valores de qualquer atributo, não apenas a classe.
  - Em geral vai produzir um conjunto muito maior de regras.
  - Limites são necessários: cobertura, acurácia são medidas de sucesso impostas para o método.

Exemplo regressão

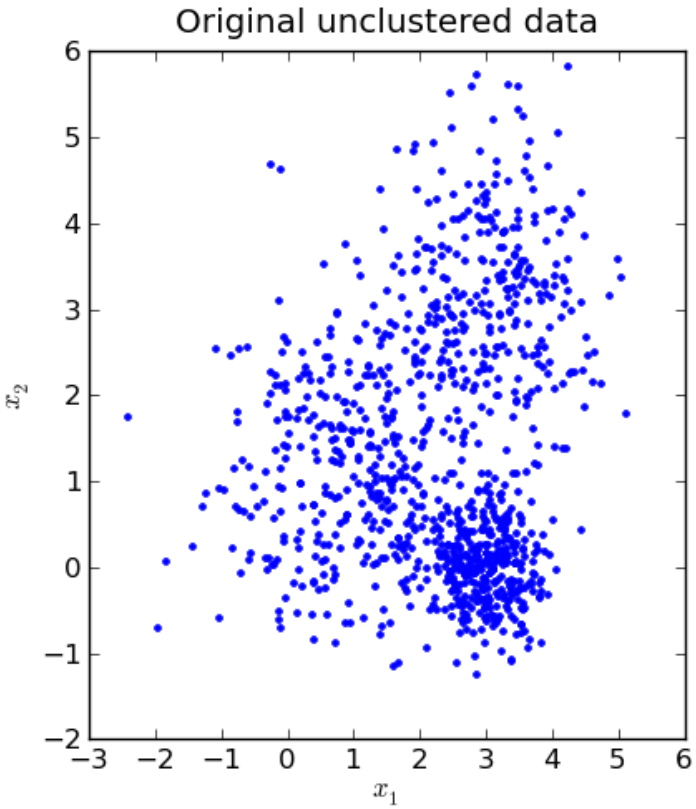
TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

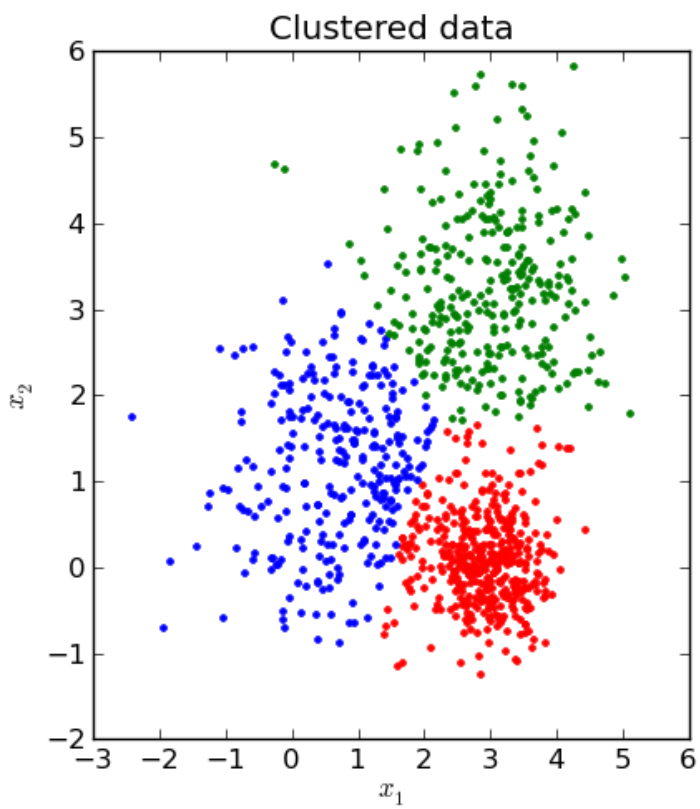
Fig: Market basket transactions

Agrupamento

- Encontrar grupos de itens que são similares em seus atributos.
- Agrupamento é aprendizado *não supervisionado*.
- Sucesso é medido subjetivamente.

Exemplo regressão

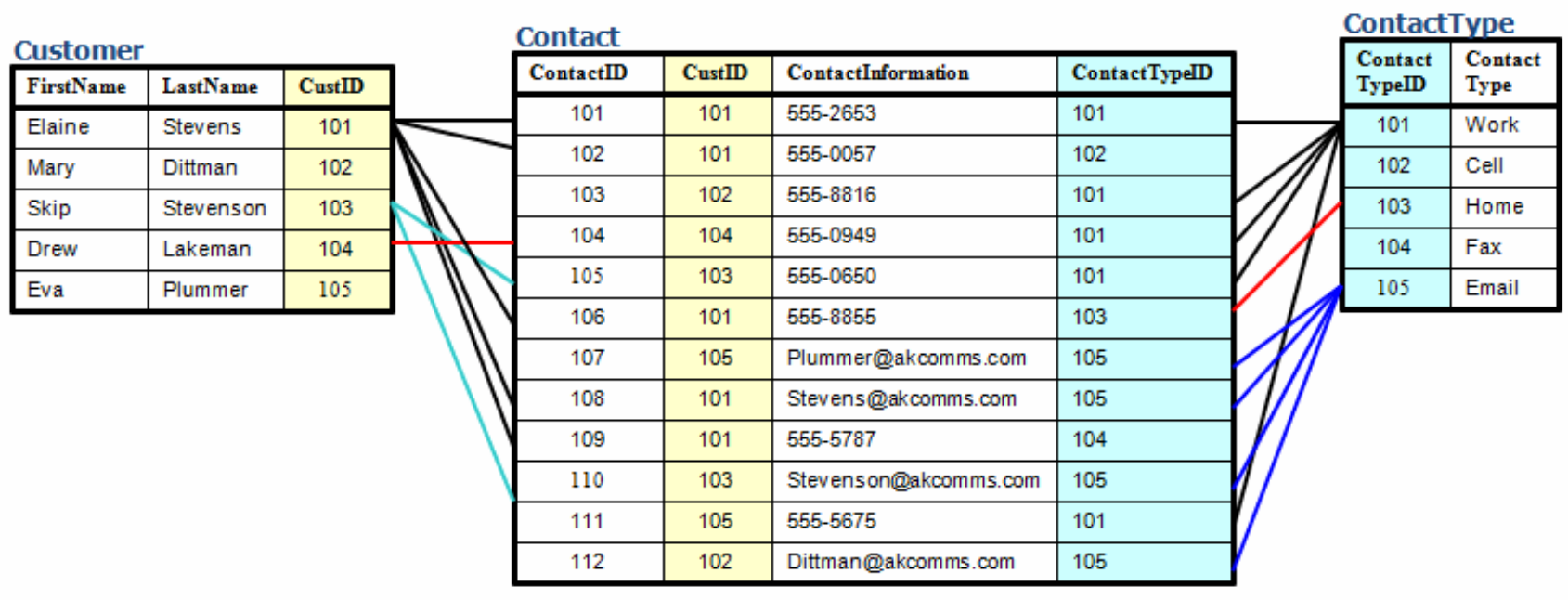




O que é um exemplo?

- Instância é o nome usado para um exemplo específico.
  - Algo a ser classificado, associado, agrupado.
  - Caracterizado por um conjunto predeterminado de atributos.
- Entrada de um esquema de aprendizado é um conjunto de instâncias (*dataset*).
- Forma mais comum de entrada.
- Não permite associação entre dados.

Dados normalizados



Customer

FirstName	LastName	CustID
Elaine	Stevens	101
Mary	Dittman	102
Skip	Stevenson	103
Drew	Lakeman	104
Eva	Plummer	105

Contact

ContactID	CustID	ContactInformation	ContactTypeID
101	101	555-2653	101
102	101	555-0057	102
103	102	555-8816	101
104	104	555-0949	101
105	103	555-0650	101
106	101	555-8855	103
107	105	Plummer@akcomms.com	105
108	101	Stevens@akcomms.com	105
109	101	555-5787	104
110	103	Stevenson@akcomms.com	105
111	105	555-5675	101
112	102	Dittman@akcomms.com	105

ContactType

Contact TypeID	Contact Type
101	Work
102	Cell
103	Home
104	Fax
105	Email

Arquivos planificados

- Um arquivo planificado é aquele em que todas as relações estão "desnormalizadas".
- Possível com qualquer número finito de relações.
- Problemático para relações com um número não especificado de elementos.
- Pode gerar um arquivo com muita redundância.
- Relações infinitas podem necessitar de procedimentos recursivos para a sua enumeração.

O que é um atributo?

- Cada instância é descrita por um conjunto pré-definido de características, chamadas atributos.
- Número de atributos pode ser diferente para cada instância. Uma solução possível é adicionar um valor "irrelevante".
- Outro problema é que a existência de um atributo pode depender do valor de um outro atributo.
- Tipos de atributos
  - Nominal
  - Ordinal
  - Intervalo
  - Valores

Atributos nominais

- Valores são símbolos distintos.
- Não há relação entre os valores.
- Apenas testes de igualdade são aplicáveis.

Atributos ordinais

- Valores são símbolos distintos.
- Há uma ordem entre os valores, mas não há ideia de distância entre eles.
- É possível compará-los com $\leq, \geq$, mas não faz sentido operar com adição ou subtração.
- Nem sempre é claro distingui-los dos atributos nominais.

Atributos em intervalos

- São quantidades, expressas em valores dentro de um conjunto.
- Exemplo: temperatura escrita em Celsius.
- Faz sentido compará-los e operá-los.

Valores

- Proporções são medidas a partir de um ponto zero.
- Exemplo: distância em metros.
- São tratadas como números reais.
- Todas as operações matemáticas são aplicáveis.

Metadados

- Algumas informações podem vir associadas aos dados fornecidos em um conjunto de instância.
- Estas informações podem ser usadas pelos algoritmos de aprendizado, mas não fazem parte de qualquer exemplo.
- Exemplos:
  - Dimensões máximas e mínimas
  - Ordenações circulares ou parciais
  - Relações de generalização, especialização, etc.

O formato ARFF

- Formato de arquivo texto.
- Seções:
  - Relações
  - Atributos (permite texto e datas)
  - Dados
- Formato permite atributos relacionais e multi-instâncias.
- Também é compatível com dados esparsos.

Problemas com a entrada

- Dados faltantes.
- Dados imprecisos.
- Dados desbalanceados.

Conhecendo seus dados

- Visualização.
- Entendendo dependências.
- Entendendo a origem e o campo.
- Coletando amostras.