Tópicos Especiais em Computação I

Pré-processamento de dados

Patrícia Lucas

Bacharelado em Sistemas de Informação
IFNMG - Campus Salinas

Salinas
Março 2021

INSTITUTO FEDERAL
Norte de Minas Gerais
Campus Salinas

Versão 1.2021

Pré-processamento de dados

Referência

Capítulo 3: Análise de dados.

Inteligência Artificial: Uma abordagem de aprendizado de máquina. Katti Faceli...[et al.]. - Rio de Janeiro: LTC, 2011.

Visão geral

Pré-processamento de dados

Apesar de algoritmos de aprendizagem de máquina serem frequentemente adotados para extrair conhecimento de conjunto de dados, seu desempenho é geralmente afetado pelo estado dos dados.

Os valores dos atributos podem apresentar diferentes características, dimensões e formatos.

Podem também apresentar ruídos e imperfeições: valores incorretos, inconsistentes, duplicados ou ausentes, etc.

Técnicas de pré-processamento de dados são frequentemente utilizadas para melhorar a qualidade dos dados por meio da eliminação ou minimização desses problemas.

Visão geral

Pré-processamento de dados

A melhoria na qualidade dos dados permitem:

- facilitar as técnicas de aprendizado de máquina;
- levar a construção de modelos mais fiéis à distribuição dos dados;
- reduzir a complexidade computacional;
- tornar o ajuste dos parâmetros do modelo mais fáceis e rápidos;
- facilitar a interpretação dos padrões extraídos pelo modelo.

Visão geral

Pré-processamento de dados

Técnicas:

- amostragem;
- tratamento para dados desbalanceados;
- limpeza;
- integração de dados;
- transformação; e
- redução da dimensionalidade.

Eliminação manual de atributos

Pré-processamento de dados

Quando um atributo não contribui para a estimativa do valor do atributo alvo, ele deve eliminado.

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Amostragem

Pré-processamento de dados

A amostragem é um processo usado na análise estatística em que um número predeterminado de observações é obtido de uma população maior.

Algoritmos de aprendizagem de máquina podem ter dificuldade em lidar com um número grande de objetos.

Eficiência computacional X acurácia (taxa de predições corretas).

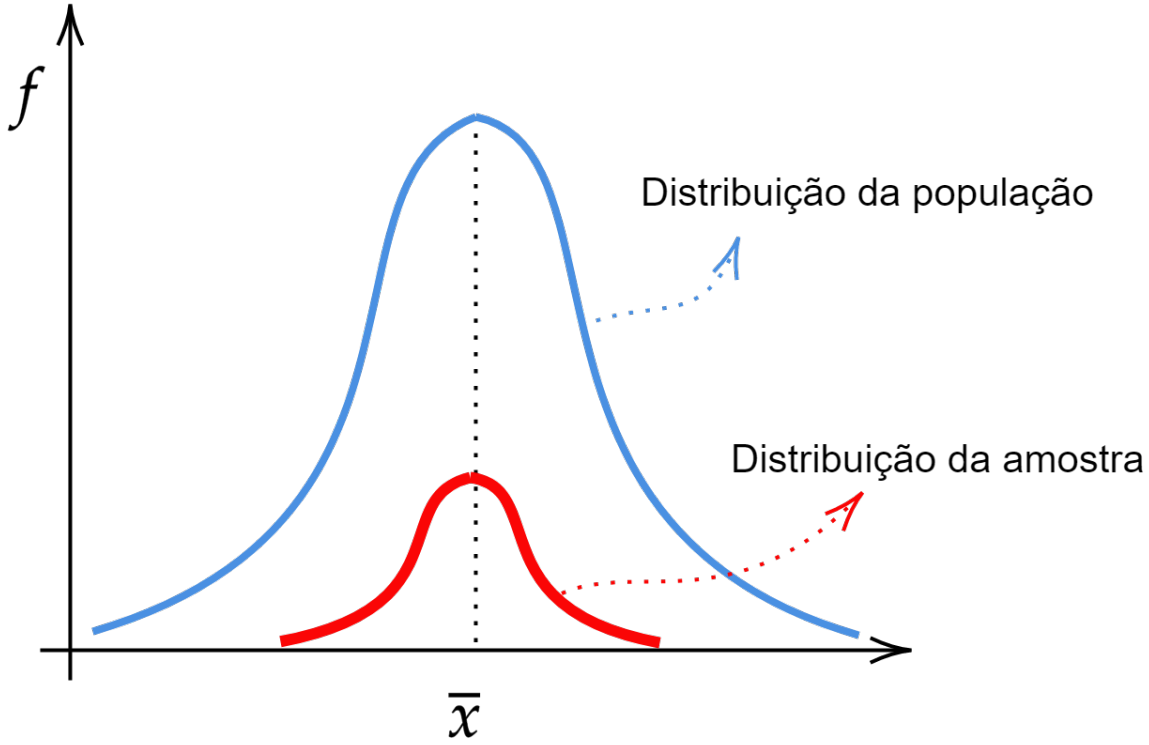
Muitas vezes, o uso de uma amostra leva ao mesmo desempenho obtido com o uso do conjunto completo, porém com um custo computacional muito menor.

Amostragem

Pré-processamento de dados

Deve ser observado que uma amostra pequena pode não representar bem o problema que se deseja modelar.

A amostra deve ser representativa do conjunto de dados original.



Amostragem

Pré-processamento de dados

Técnicas de amostragem:

- Amostragem aleatória simples.
- Amostragem estratificada: usada em problemas de classificação.
  - Manter o mesmo número de objetos em ambas as classes.
  - Manter a proporção do número de objetos do conjunto original.
- Amostragem progressiva: começa com uma amostra pequena e aumenta progressivamente enquanto a acurácia continuar a melhorar.

Dados desbalanceados

Pré-processamento de dados

O problema de **dados desbalanceados** é tópico da área de classificação de dados.

Esse problema é comum em aplicações em que dados de um subconjunto das classes aparecem com uma frequência maior que os dados das demais classes.

Exemplo: supondo que 80% dos pacientes que vão a um determinado hospital estão doentes, seu conjunto de dados apresentará então 20% de seus objetos relacionados a pacientes saudáveis.

Classe majoritária: pacientes doentes.
Classe minoritária: pacientes saudáveis.

Dados desbalanceados

Pré-processamento de dados

Quando algoritmos são alimentados com dados desbalanceados, eles tendem a favorecer a classificação de novos dados na classe majoritária.

Caso o balanceamento das classes não seja possível, o balanceamento artificial, através da criação de **amostras sintéticas**, pode ser usado.

Limpeza de dados

Pré-processamento de dados

Conjuntos de dados podem apresentar dificuldades relacionadas à qualidade dos dados:

- dados ruidosos: que possuem erros ou valores discrepantes.
- inconsistentes: que contradizem valores de outros atributos do mesmo objeto.
- redundantes: dois ou mais objetos com mesmos valores em todos os atributos.
- incompletos: ausência de valores.

Essas deficiências podem ser causadas por: problemas em equipamentos que realizam coleta, transmissão e armazenamento dos dados ou por erro humano.

Limpeza de dados

Pré-processamento de dados

Dados incompletos:

- Eliminar o objeto com valores ausentes.
- Utilizar média, moda ou mediana dos valores conhecidos do atributo.

  - Ex: usar a média do atributo de uma classe.

  - Ex: para dados que possuem relação temporal, a medida pode ser calculada usando os objetos associados ao instante imediatamente anterior e posterior ao objeto modificado.
- usar um modelo para estimar o valor do atributo.

Dados inconsistentes e redundantes:

- Exclusão.

Ruídos podem ser dados inconsistentes ou *outliers*!

Conversão Simbólico-Numérico

Quando o atributo é nominal e assume apenas dois valores, se denotam a presença ou ausência de uma característica ou se apresentam uma relação de ordem, um dígito binário é suficiente.

Exemplo: (doente = 0 | saudável = 1) ou (frio = 0 | quente = 1).

Conversão Simbólico-Numérico

Para atributos com mais de dois valores:

- Se não houver uma relação de ordem, a inexistência também dever continuar após a transformação.
- Se existe a relação de ordem, a codificação deve preservá-la.

Valor ordinal	Código cinza	Código termômetro
Primeiro	000	00000
Segundo	001	00001
Terceiro	011	00011
Quarto	010	00111
Quinto	110	01111
Sexto	100	11111

Conversão Numérico-Simbólico

- Se o atributo for do tipo discreto e binário, com apenas dois valores, basta associar um nome a cada valor.
- Se o atributo for formado por uma sequência de bits sem uma relação de ordem, cada sequência pode ser substituída por um nome.
- Nos demais casos, usar um método de discretização que melhor represente o domínio do problema.

Transformação nos dados

Pré-processamento de dados

Atributos numéricos: algumas vezes um atributo numérico precisa ser transformado em outro valor numérico. Isso ocorre quando:

- Os limites inferior e superior de valores dos atributos são muito diferentes.
- Vários atributos estão em escalas diferentes.

Essa transformação é realizada para evitar que um atributo predomine sobre o outro.

Transformação nos dados

Normalização por reescala: defini uma nova escala de valores, limite mínimo e máximo, para todos os atributos.

Exemplo de normalização min-max:

$$v_{novo} = min + \frac{v_{atual} - menor}{maior - menor}(max - min) \quad (1)$$

*Para limite superior 1 e inferior 0, max = 1 e min = 0.

Transformação nos dados

Pré-processamento de dados

Normalização por padronização: defini um valor central e um valor de espalhamento comuns para todos os atributos.

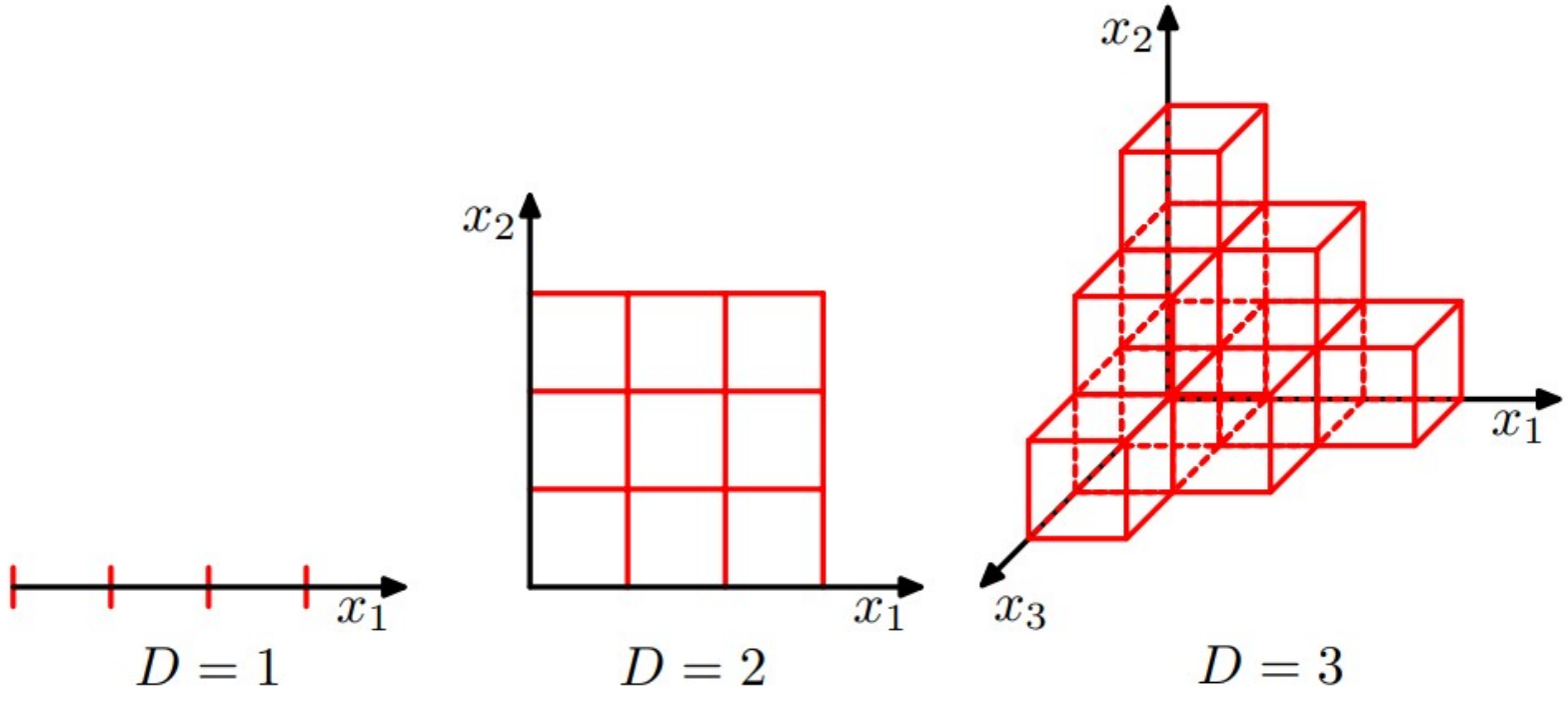
Exemplo de padronização dos valores com média (μ) = 0 e variância (σ) = 1:

$$v_{novo} = \frac{v_{atual} - \mu}{maior - \sigma} \tag{2}$$

Redução de dimensionalidade

Pré-processamento de dados

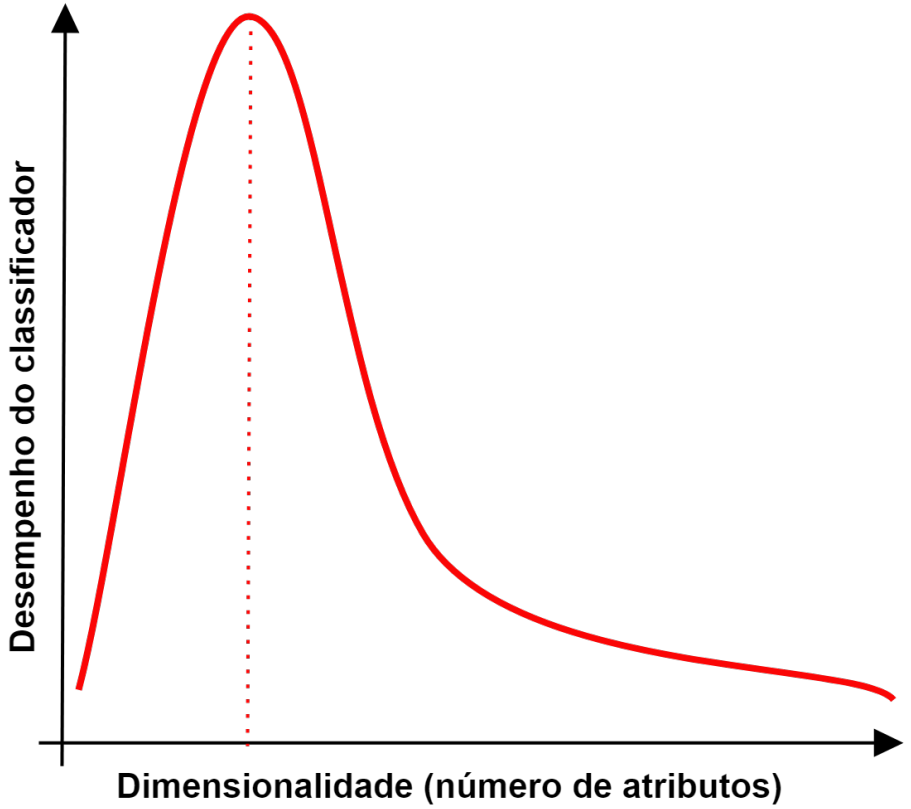
Maldição da dimensionalidade: se cada atributo é visto como uma coordenada em um espaço d-dimensional, em que d é o número de atributos, o hipervolume que representa esse espaço cresce exponencialmente.



Redução de dimensionalidade

Pré-processamento de dados

Na prática, a maldição da dimensionalidade implica que para um dado tamanho de amostras, existe um número máximo de características a partir do qual o desempenho do classificador irá degradar, ao invés de melhorar.



Redução de dimensionalidade

Técnicas de redução de dimensionalidade:

- *Agregação:* substituem os atributos originais por novos atributos formados pela combinação de grupos de atributos e consequentemente resulta em perda de informação.

  - Exemplo: análise de componentes principais (PCA).

- *Seleção:* mantêm uma parte dos atributos originais e descartam os demais atributos. Essas técnicas procuram um subconjunto ótimo de atributos de acordo com um determinado critério.

  - Exemplo: embutida, baseada em filtro ou baseada em *wrapper*.