Redução de Dados

Agente Educacional
Sérgio M. Dias

Agenda

Redução de dados
Maldição da dimensionalidade
Estratégias para redução de dados

Objetivos

Redução do custo computacional

(Bases de dados podem ser muito volumosas)

Eliminação de redundância

Evitar maldição da dimensionalidade

Maldição da Dimensionalidade

Em geral, na busca por melhores resultados na aprendizagem, criam-se vetores de características com mais informações.

Redução de Dados

Maldição da Dimensionalidade

Para ilustrar o problema, vamos considerar o exemplo da figura a seguir, com duas características x 6 e x 7 e três classes.

Para classificar o padrão x, dividimos o espaço em células de tamanho igual e atribuímos a x a classe mais frequente dentro da célula.

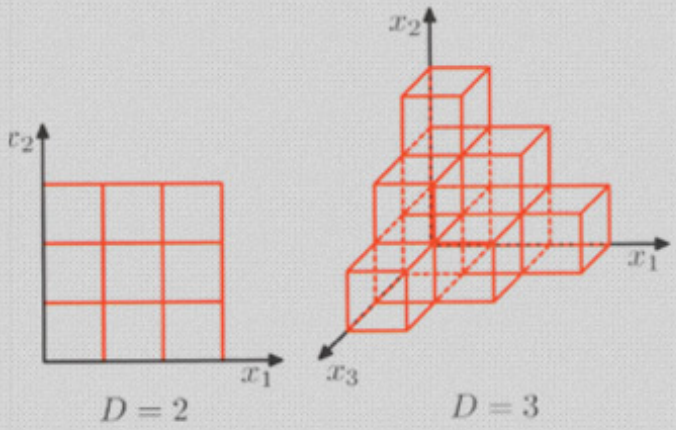
Redução de Dados

Maldição da Dimensionalidade

Conforme adicionamos características ao vetor, o número de células cresce de forma exponencial.

Em função disso, quanto mais características temos, mais base de dados precisamos ter para preencher todas as células.

Na maioria dos problemas, entretanto, a quantidade de dados disponível para a aprendizagem é limitada.



Redução de Dados

Maldição da Dimensionalidade

“Precisamos evitar o excesso de atributos”

Estratégias

Redução de dimensionalidade
Seleção de atributos
Amostragem
Compressão

Redução de dimensionalidade

Dimensões

Número de atributos

Eliminar atributos pouco relevantes

Reduzir tempo de processamento

Facilitar visualização

Quanto mais dimensões, mais difícil é visualizar dados

Em um espaço cartesiano, só vemos bem até 3 dimensões

Redução de Dados

Redução de dimensionalidade

Antes de tudo, remover identificadores

- Quase sempre irrelevantes, quando não atrapalham a análise de dados

Técnicas para redução de dimensionalidade

- PCA (*Principal Components Analysis*)
- Seleção de atributos
  - Remoção de atributos irrelevantes
  - Pode ser feita usando matriz de correlação

Matriz de correlação

Analitos	Precipitação	Acetato	Cloreto	Nitrito	Brometo	Nitrato	Fosfato	Sulfato
Acetato	**-1**	1						
Cloreto	-0,05	0,42	1					
Nitrito	**-1**	0,03	-0,02	1				
Brometo	-0,1	0,32	0,32	**0,76**	1			
Nitrato	**-1**	**0,76**	**0,83**	-0,12	0,46	1		
Fosfato	-0,33	0,34	0,44	-0,04	0,21	**0,56**	1	
Sulfato	0,05	0,42	**0,93**	-0,06	**0,63**	**0,94**	**0,55**	1

Seleção de atributos

Manual

Quando há conhecimento profundo sobre os dados,
normalmente com a participação de um especialista no negócio

Automático

Entre variáveis contínuas, pode-se usar correlação

Exercício

Seleção manual de atributos
Selecionar atributos por limiar de correlação

Obrigado!

Agente Educacional

Sérgio Mariano Dias

sergio.dias@serpro.gov.br | #31 6539

Demais agentes educacionais sobre o assunto

Marcelo Pita | marcelo.pita@serpro.gov.br | #81 8794

Gustavo Torres | gustavo.gamatorres@serpro.gov.br | #31 6950