Sistemas Inteligentes Aplicados:
Pré-processamento dos dados – Operações

Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Pré-processamento

- Selecionar os dados é uma das etapas mais importantes do processo:
  - Em alguns casos um bom pré-processamento é mais importante que a escolha do algoritmo de aprendizado de máquina
  - O sucesso da mineração depende da qualidade dos dados escolhidos
  - A maioria dos dados disponíveis não foram inicialmente concebidos para aprendizado de máquina

Pré-processamento

- Operações de pré-processamento
  - Tratamento de ruídos (a seguir)
  - Seleção de informações
  - Transformações nos dados
  - Balanceamento de dados e amostragem

Ruídos

- Ruídos são basicamente **erros** presentes nos dados
  - Erros de digitação
  - Erros de captura de sensores automáticos (ex.: artefatos em imagens)
  - Problemas no próprio pré-processamento (ex.: detecção incorreta de contornos em imagens)
  - Fatores externos que não podemos controlar (ex.: usuário decidiu informar sua idade incorretamente)

Ruídos (2)

- Normalmente, não é possível eliminar todos os ruídos e precisamos conviver com eles
- Os algoritmos tem boa tolerância a ruídos, desde que sejam pouco frequentes

Exemplos de Ruídos

- Valores inconsistentes (ex.: idade negativa)
- Duplicações nos dados
- Outliers (ou pontos discrepantes ou anomalias):
  - Valores muito diferentes dos demais
  - Ex.: pessoa centenária na base
  - Nem sempre são ruídos

Seleção de informações

- Seleção de informações
  - Descartar atributos irrelevantes (ex.: ids)
  - Descartar instâncias (ex.: duplicadas, com ausências)

Transformações nos dados

- Nivelar inconsistências (usar idade mínima onde for negativa)
- Estimar valores ausentes: média, mediana, moda, etc
- Converter escalas (ex.: cidades para estados; valores numéricos entre 0 a 1; …)
- Criar/mesclar atributos (ex.: IMC combina altura e peso no atributo)

Balanceamento

- Grupos de instâncias desbalanceados: natural ao domínio ou algum problema na coleta
- Muitos algoritmos ficam viciados/tendenciosos em dados desbalanceados
  - Favorecem a classe majoritária (mais comum)
- Estratégia 1: **eliminar instâncias** (via amostragem estratificada)
- Estratégia 2: **repetir instâncias** pouco frequentes

Balanceamento (2)

- Por que repetir instâncias?
  - Alguns grupos de instâncias são muito pouco frequentes
  - Balancear por eliminação compromete a **representatividade**

Balanceamento (3)

Thanks to machine-learning algorithms,
the robot apocalypse was short-lived.

Amostragem de instâncias

- Pode ser usada para investigação preliminar e para análise final dos dados
- Diminui tempo de processamento:
  - Possibilita uso de algoritmos mais avançados
  - Alternativa a processamentos custosos

Amostragem de instâncias (2)

- Consiste em extrair um subconjunto da população de dados (ou universo de dados)
- Se bem aplicada, preserva ou melhora os resultados
  - Ex.: em classificação, amostra com classes balanceadas melhora resultados

Amostragem de instâncias (3)

- A amostra resultante precisar ser **representativa**:
  - Deve preservar as propriedades estatísticas do população do qual a amostra foi extraída
  - Média da população $\cong$ média da amostra;
    Desvio da população $\cong$ desvio da amostra;
    etc
  - Deve permitir tirar conclusões sobre o todo a partir de uma parte (indução!)

Amostragem de instâncias (4)

- Tipos de amostragem
  - **Aleatória simples**
  - **Estratificada**
  - **Progressiva**

Amostragem simples

- **Amostragem aleatória simples**: sortear elementos aleatoriamente
- Variações
  - Com reposição
  - Sem reposição
  - São semelhantes quando amostra é bem menor que população

Amostragem estratificada

- **Amostragem estratificada**: leva em conta grupos/classes de instâncias na população
  - Ex.: número de jogadores de basquete e halterofilistas ao predizer esporte do atleta
- Variações:
  - **Balanceada**: iguala mesmo número de grupos na amostra
  - **Proporcional**: número de grupos na amostra reflete a população

Amostragem progressiva

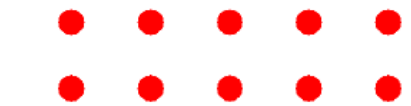
- **Amostragem progressiva**: começa com uma amostra pequena e vai aumentando o número de elementos gradualmente
  - Algoritmo de aprendizado é rodado várias vezes
  - Processo continua até estabilizar (resultados obtidos param de melhorar)

Amostragem progressiva (2)

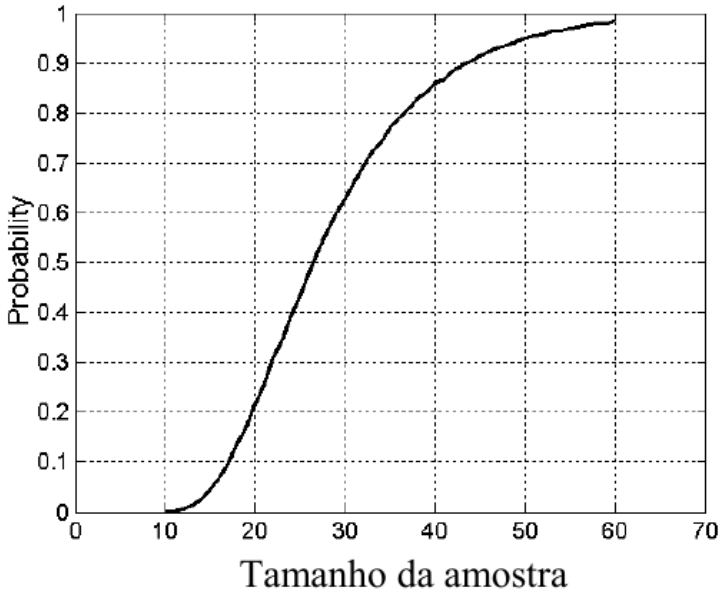
- Amostragem progressiva ajuda a determinar o melhor tamanho para a amostra
  - **Muito grande**: amostra representativa, mas custo elevado
  - **Muito pequena**: custo baixo, mas resultados ficam ruins

Amostragem progressiva (3)

- Determinando um bom tamanho
  - Qual o tamanho ideal para obter 10 grupos de elementos da amostra de acordo com o tamanho da amostra?

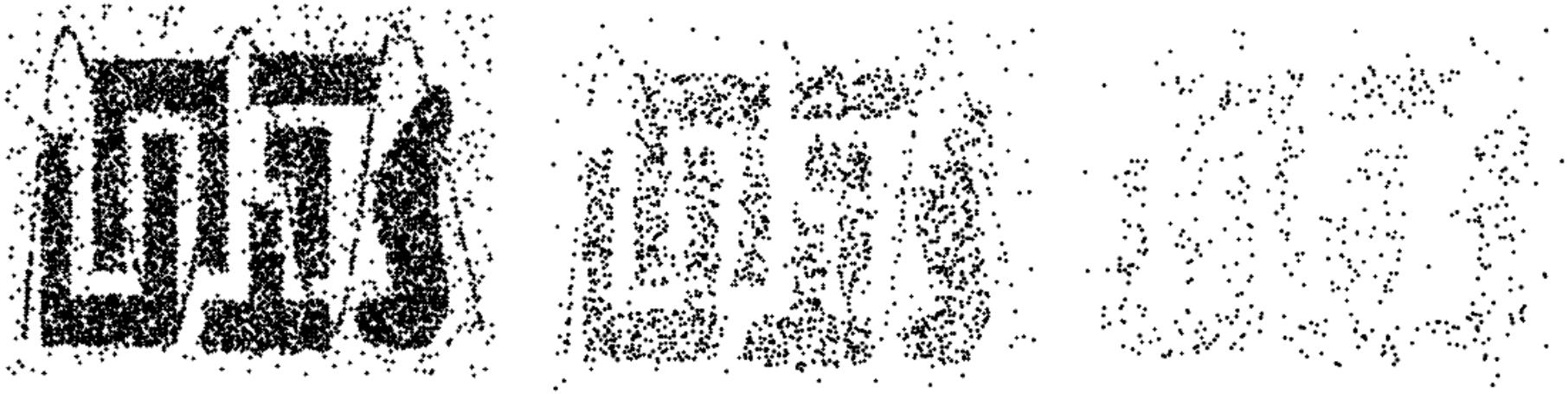


10 grupos de pontos



Amostragem progressiva (4)

- Importância do tamanho da amostra: exemplo para **agrupamento** de dados

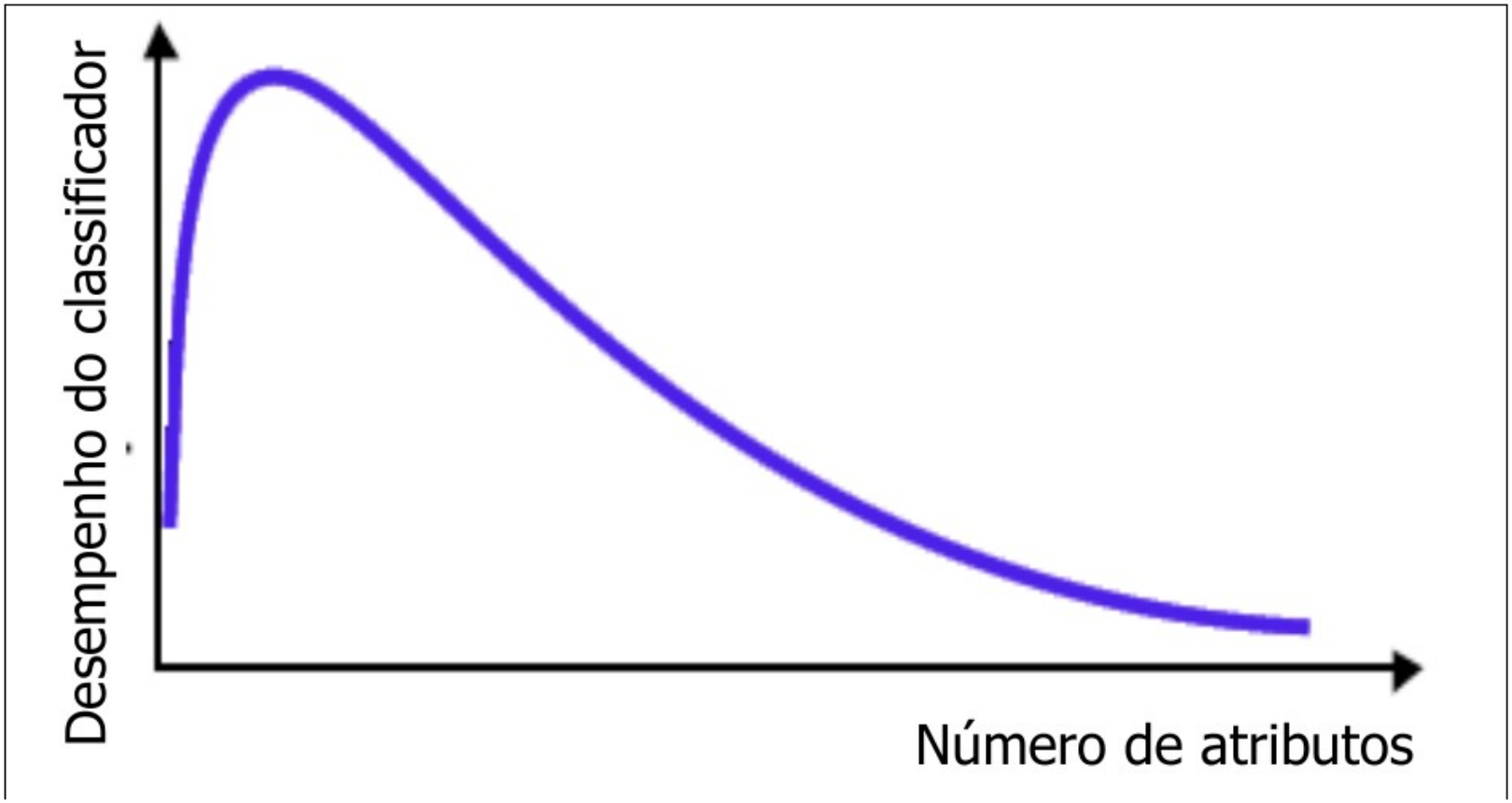


8000 pontos | 2000 Pontos | 500 Pontos

Maldição da dimensionalidade

- Excesso de atributos é causa a **maldição da dimensionalidade** (curse of dimensionality)
- Excesso de atributos confunde mais do que ajuda

Maldição da dimensionalidade (2)

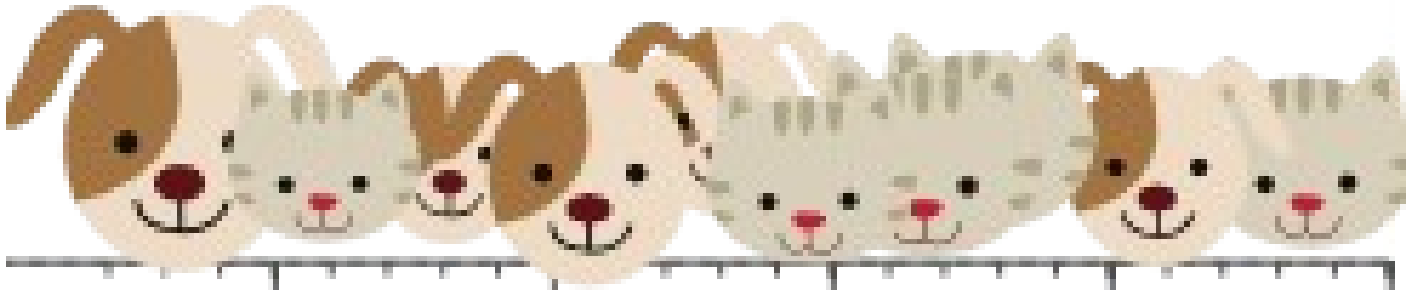


Maldição da dimensionalidade (3)

- Exemplo:
  - Separar cães de gatos
  - Treinamento com 10 instâncias
  - Atributos fictícios com 5 possíveis valores cada

Maldição da dimensionalidade (4)

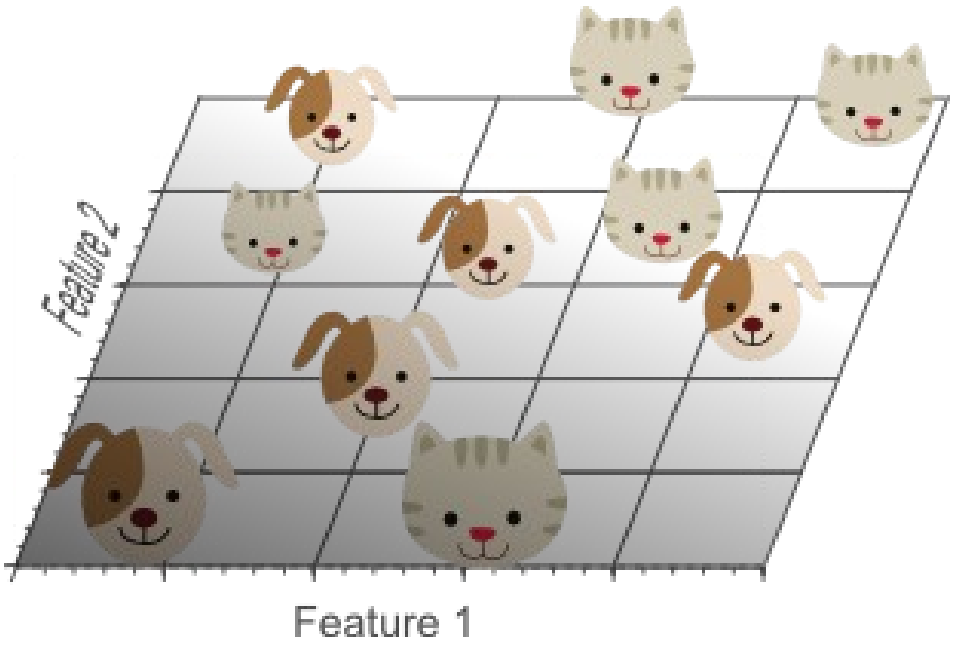
- Com 1 atributo, existem 5 posições válidas
- As 10 instâncias cobrem (com repetição) 100% das posições



Feature 1

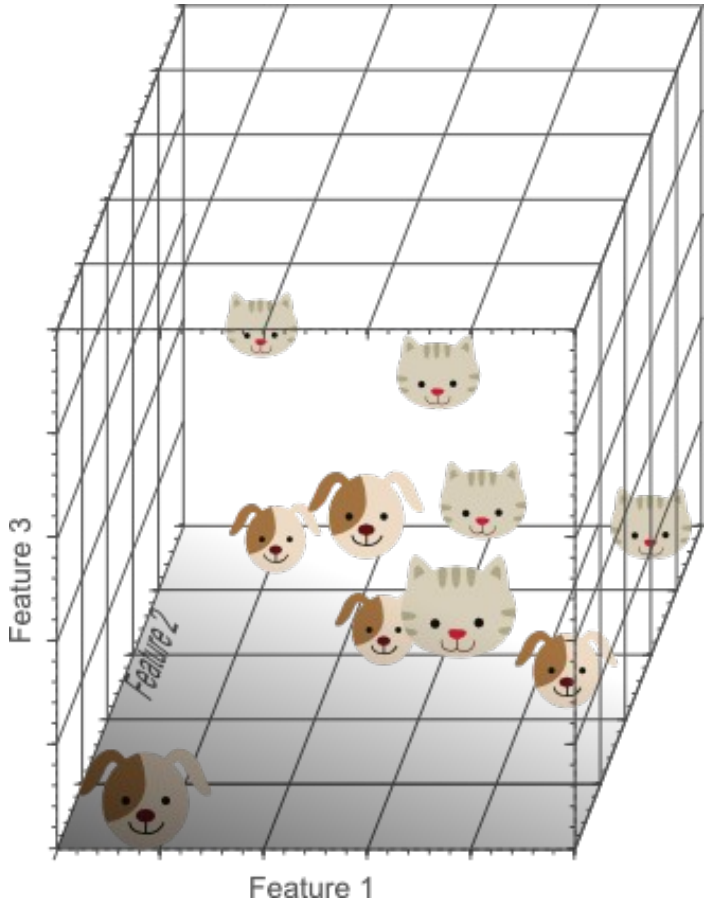
Maldição da dimensionalidade (5)

- Com 2 atributos de 5 valores cada, existem 25 posições válidas
- As 10 instâncias cobrem 40% das posições



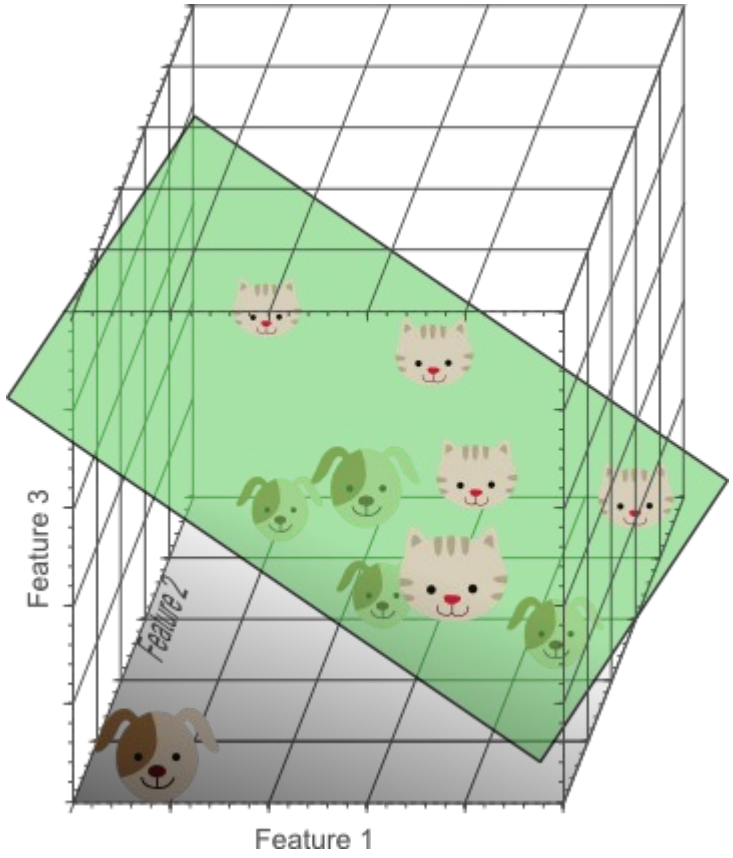
Maldição da dimensionalidade (6)

- Com 3 atributos, apenas 8% das posições são cobertas (10/125)

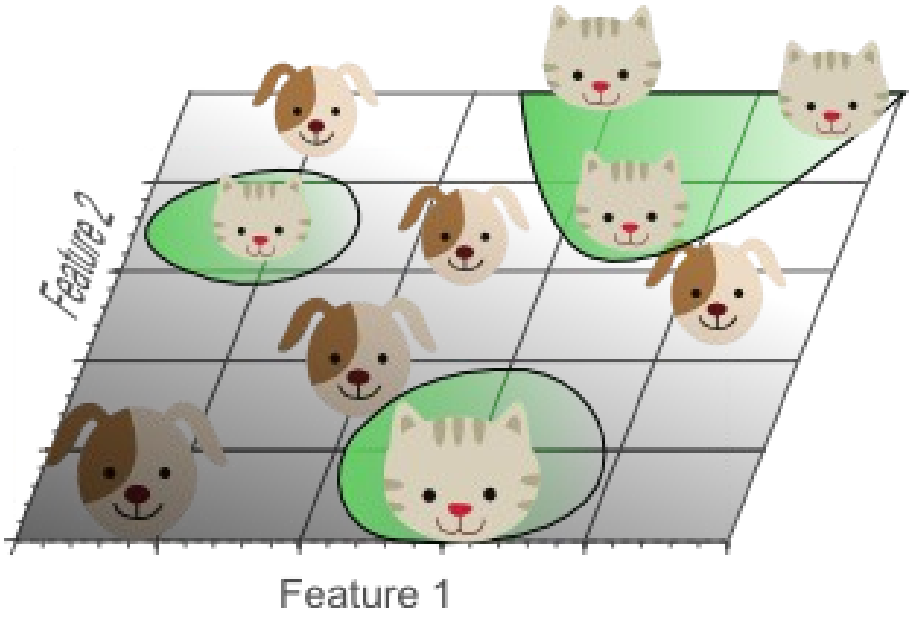


Maldição da dimensionalidade (7)

- No R^3, parece uma haver uma boa divisão



- Projetando a divisão no R^2, parece overfitting

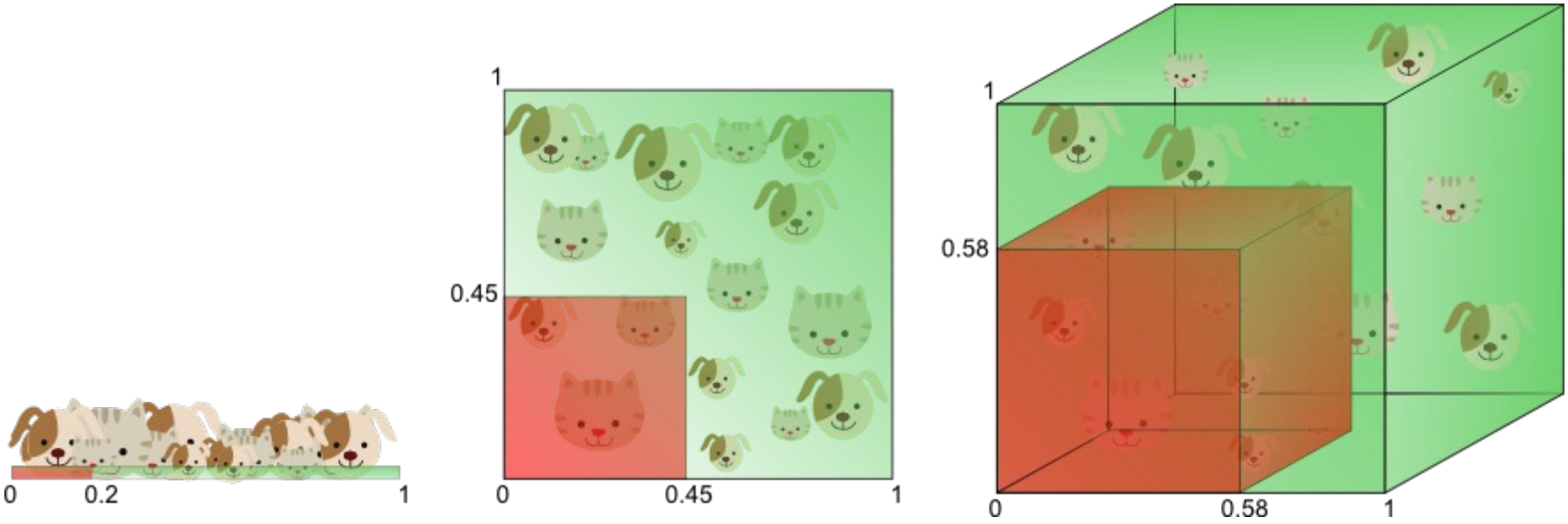


Maldição da dimensionalidade (8)

- Mais atributos → mais generalizações ruins, **mas que parecem boas**
- O número de atributos cresce suavemente
  - Mas o número instâncias necessárias para uma boa generalização cresce exponencialmente

Maldição da dimensionalidade (9)

- Outra analogia: qual porcentagem do espaço é necessário cobrir para obter-se 20% das instâncias?

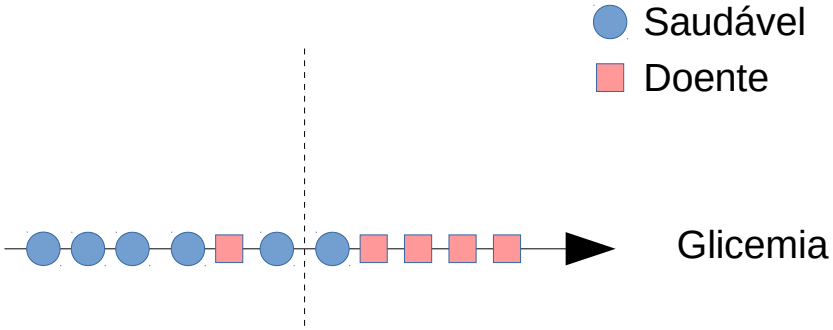


Maldição da dimensionalidade (10)

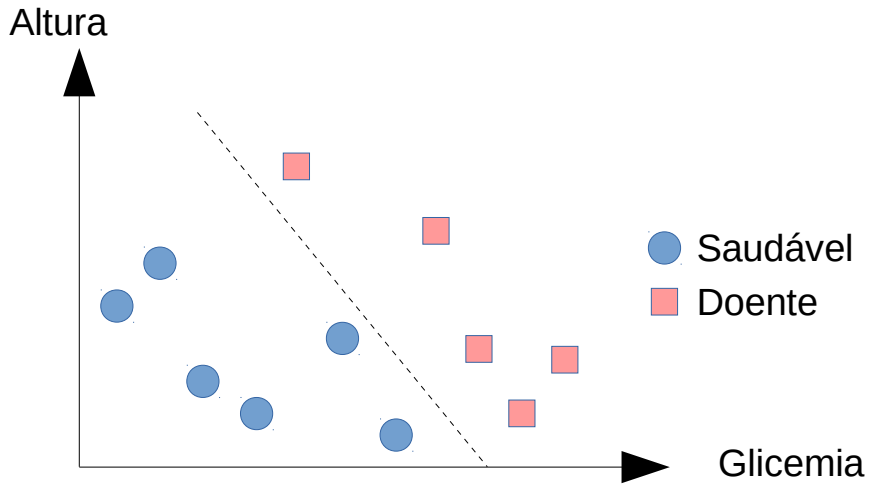
- A distância entre as instâncias aumenta com o acréscimo de atributos
  - Inclusive para instâncias da mesma classe
  - Viola a regra informal: “coisas parecidas andam juntas”
- O problema é agravado quando existem atributos irrelevantes (próximo slide)

Maldição da dimensionalidade (11)

- Glicemia ajuda a detectar diabetes



- Altura não ajuda



Maldição da dimensionalidade (12)

- Alguns doentes, por coincidência, eram mais altos que alguns pacientes saudáveis
- O modelo do R^2 acabou aprendendo que pessoas altas são mais propensas a diabetes (errado!)
- Datasets disponíveis para AM geralmente são pequenos e não cobrem todos os casos
  - Cada atributo acrescentado dá mais chance ao acaso para agir

Reduzindo a dimensionalidade

- Estratégias: **agregação**, **extração de características**, **combinação de atributos numéricos**, **seleção automática de atributos**
  - Melhoram resultados de alguns algoritmos
  - Diminuem memória e tempo de processamento
  - Simplificam a interpretação do conhecimento gerado e a visualização dos dados
  - Eliminam atributos irrelevantes e reduzem ruídos

Agregação

- Combinar atributos parecidos em um único
  - Exemplo: IMC combina altura e peso
  - Em text mining: stemming junta várias palavras parecidas (faz, fazer, fazendo, …) em um único atributo

Agregação (2)

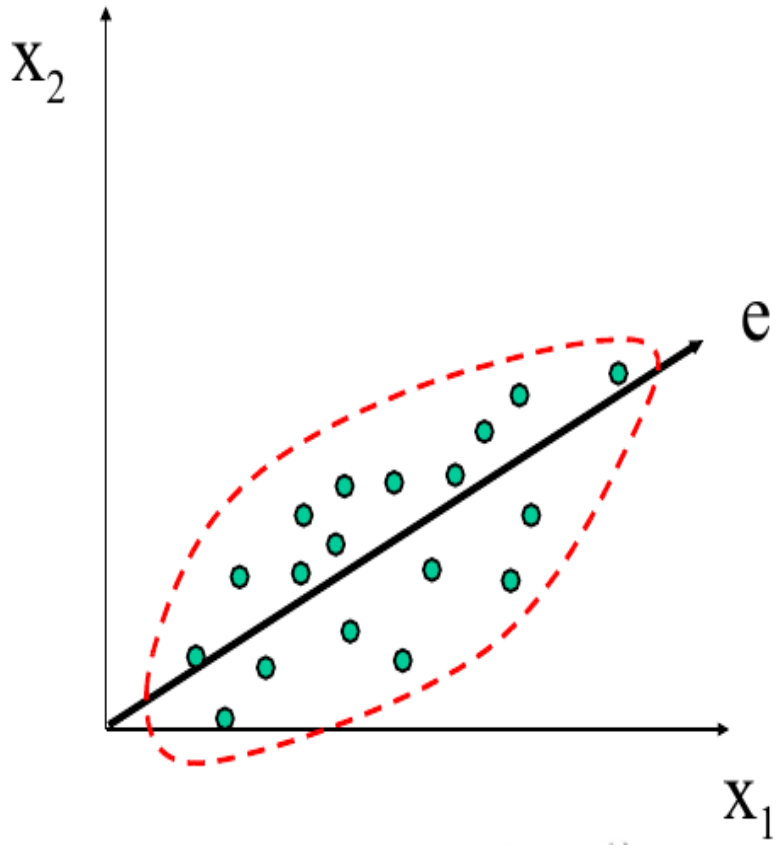
- Estabiliza dados: causa menor variabilidade nos dados
- Agregação assistida por máquina para atributos numéricos
  - Feita para reter a maior parte da informação
  - Exemplo de método: PCA

Agregação (3)

- Combinação de atributos: agregação assistida por máquina para atributos numéricos
  - Feita de forma estratégica para reter a maior parte da informação
  - Exemplo de método: PCA

Agregação (4)

- PCA: método estatístico para compressão de dados
- Exemplo, reduzir os atributos x_1 e x_2 para o atributo ***e***
- Projeção ***e*** deve tentar preservar variação nos dados (por exemplo, proximidade e ordem)



Extração de características

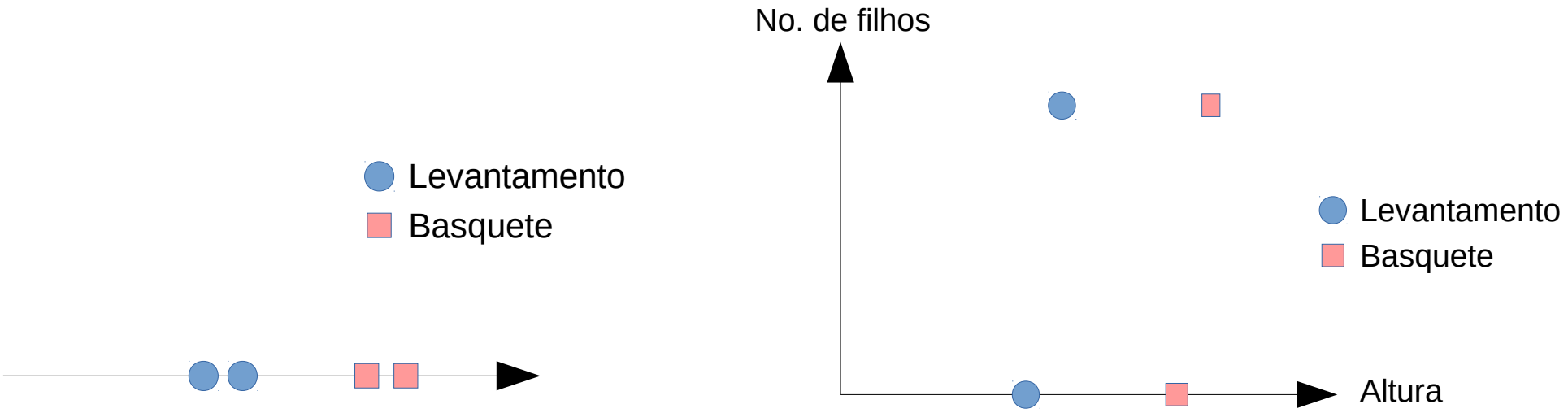
- **Extração de características** (ou de traços) são versões refinadas de atributos mais brutos
- Exemplo: reconhecimento de faces
  - Imagens são grandes matrizes onde cada célula (pixel) pode ser um atributo
  - 1024x768 pixels em escala de cinza→ 786.432 atributos!
  - Representação vetorial (contornos) tem número bem menor de informações

Seleção de atributos

- Objetivo: **descartar atributos desnecessários**
  - Atributos irrelevantes (ex.: nome do paciente em diagnóstico de doenças; atributos constantes)
  - Atributos redundantes (ex.: base com preço do produto, porcentagem do desconto e valor do desconto)
  - Stop-words em mineração de texto: não adicionam informação relevante ao processamento

Seleção de atributos (2)

- Quando atributos irrelevantes estão presentes, a maldição da dimensionalidade é ainda mais acentuada:



Seleção de atributos (3)

- No exemplo, com uma ajudinha do acaso, o atributo “número de filhos” deixou as instâncias longes de seus pares e próximas da outra classe
- A regra informal “coisas parecidas andam juntas” foi violada

Seleção de atributos (4)

- Seleção automática de atributos: consiste em usar técnicas estatísticas para auxiliar o aprendizado de máquina. Permite
  - Identificar atributos importantes
  - Obter melhores resultados
  - Minimizar efeitos de ruídos
  - Reduzir custo da coleta de dados

Seleção de atributos (5)

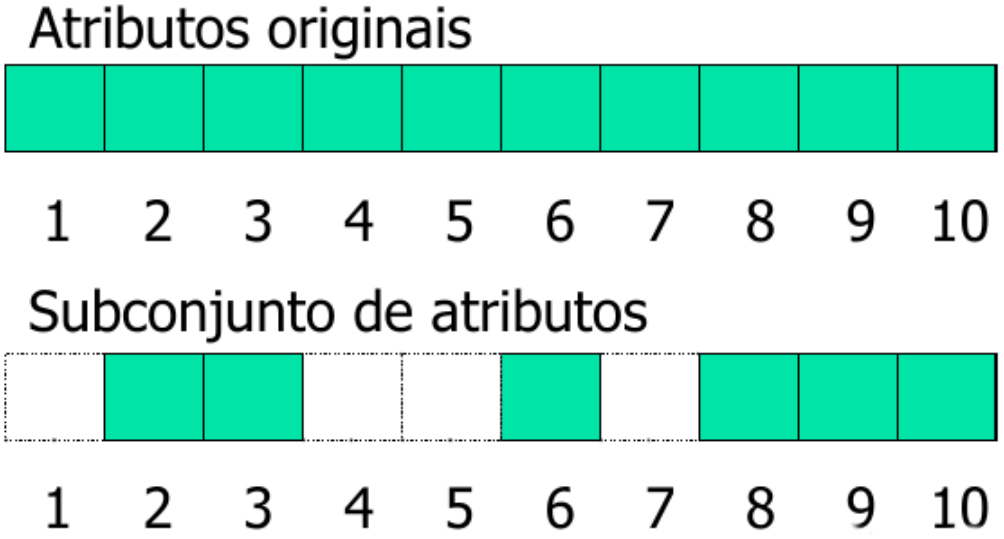
- **Impacto**:
  - **Melhora os resultados**: quando existem atributos que não são bons ou quando a dimensionalidade é muito alta
  - **Mantém os resultados**: quando alguns atributos são redundantes (calculados em função de outros)
  - **Piora os resultados**: quando um atributo importante é removido
    - Muitas vezes vale a pena, pois barateia a coleta

Seleção de atributos (6)

- Estratégias: por ordenação ou por subconjunto
- **Seleção por ordenação**: identifica individualmente os melhores atributos

Seleção de atributos (7)

- **Seleção por subconjunto**: busca identificar subconjuntos cujos atributos sejam mais relevantes quando usados juntos



Seleção de atributos (8)

- Exercício: ordenar os atributos A1..A5 dos mais importantes para os menos importantes para o diagnóstico

A1	A2	A3	A4	A5	Diagnóstico
0	0	1	1	1	0
1	1	1	1	0	1
0	0	1	0	0	1
1	0	0	0	0	0
0	1	0	1	0	1
1	0	0	0	0	1

Seleção de atributos (9)

- Estratégias de seleção automática de atributos
  - **Filtros**: processo independente utilizado antes do aprendizado de máquina
  - **Wrappers** (embrulhos): busca melhores atributos para um algoritmo de aprendizado de máquina específico
  - **Embebbed** (embarcado): interno ao algoritmo de aprendizado de máquina

Seleção de atributos (10)

- **Filtros**
  - Fácil e rápido, mas resultados podem ser piores que das outras estratégias
  - Medidas usadas: informação mutua ou correlação de atributos
  - No Weka: InfoGain (seleção por ordenação) e CfsSubsetEval (seleção por subconjunto)

Seleção de atributos (11)

- **Wrappers**: visão geral
  - Rodar algoritmo de aprendizado de máquina várias vezes
  - Variar atributos em cada rodada
  - Selecionar aqueles atributos que levaram que maximizam o desempenho do algoritmo aprendizado de máquina
  - Pode ser custoso

Seleção de atributos (12)

- **Embebbed**:
  - Faz parte do algoritmo de aprendizado de máquina
  - Exemplo clássico: árvores de decisão usam a medida **ganho de informação** para cada atributo

Exercícios

- 1. Cite e compare as duas grandes abordagens para balancear os dados.
- 2. Explique porque a inserção de atributos extras nem sempre melhora o aprendizado de máquina.
- 3. Compare seleção de atributos por ordenação e por subconjunto.

Exercícios (2)

- 4. Uma unidade básica de saúde recebe mensalmente 480 pacientes em média. No último mês, foram atendidos 360 pacientes resfriados, 96 pacientes com virose e 24 com ferimentos leves. Indique quantos exemplos de pacientes por grupo existem em:
- (a) uma amostra estratificada balanceada de 18 pacientes
- (b) uma amostra estratificada proporcional de 20 pacientes
- Faça as aproximações que julgar necessárias

Exercícios (3)

- Abrir weka para exercícios a seguir
- 5. Aplique balanceamento por repetição e eliminação em uma das bases (resample)
- 6. Faça seleção de atributos na base Iris: (a) por ordenação; (b) por subconjunto; (c) PCA; (d) por wrapper; (e) embebbed

Pontos chaves

- Seleção de atributos
- Escalas de dados
- Conversão entre escalas
- Amostragem, balanceamento e agregação

Agradecimentos/referências

- Notas de aula do Prof. André de Carvalho (USP)