Introdução a Ciências de Dados

Aula 3: Tratamento e transformação de dados

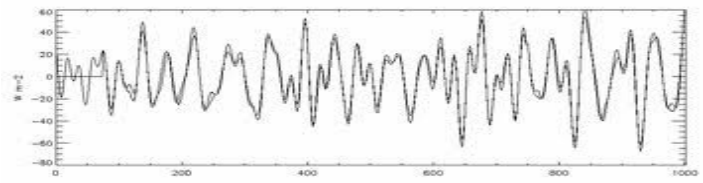
Francisco A. Rodrigues
ICMC/USP
francisco@icmc.usp.br

Aula 3: Tratamento e transformação de dados

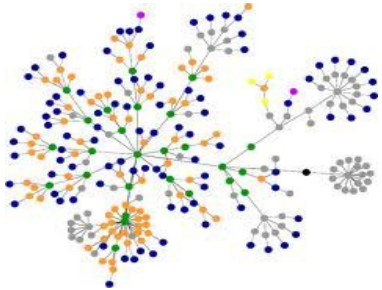
- Técnicas para Tratamento e Transformação de Dados.
- Normalização
- Análise dos componentes principais.

Tipos de Dados

Dados podem ter diferentes formatos:



Séries temporais



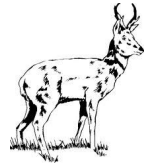
Grafos

Textos

Áudios

Vídeos

Páginas web



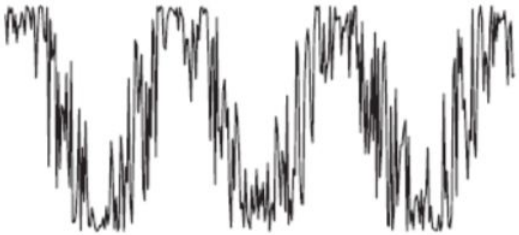
Imagens

Geralmente transformados para o formato atributo-valor

Dados com problemas

Ruídos

(a) Time series.



(b) Time series with noise.

Dados com problemas

Amostragem

(a) 8000 points (b) 2000 points (c) 500 points

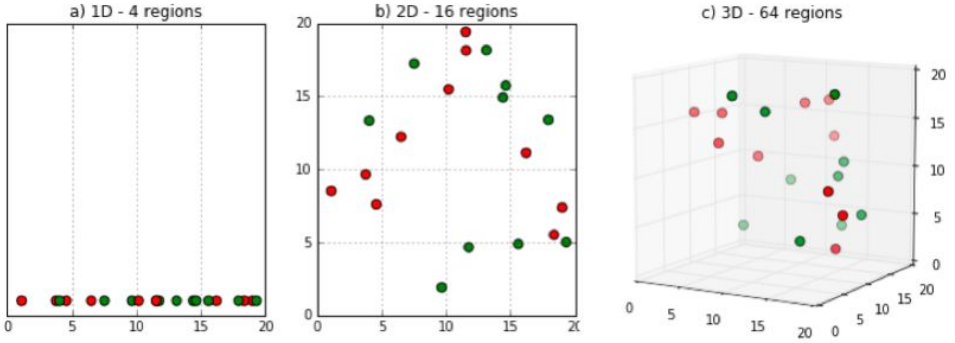
Dados com problemas

Dados incompletos

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre
100	João	SP	22	M	90,5	Baixa
102	Maria	SP	39	F	?	Baixa
230	Rubens	?	?	M	120,2	?
40	?	RS	21	M	85,9	?
543	Marta	?	40	F	88,7	Alta
12	Ana	AM	32	F	?	Média
130	?	SP	29	M	?	Média
123	José	RJ	30	M	80,1	Baixa

Maldição da dimensionalidade

- Se dividirmos uma região do espaço em células regulares, o número células cresce exponencialmente com a dimensão do espaço.
- Assim, o número de amostras deve crescer exponencialmente para garantir que nenhuma célula fique vazia.
- O poder preditivo de um método aumenta com o número de atributos, mas depois diminui (fenômeno Hughes).



Dados com problemas

- Dados correlacionados;
- Dados duplicados;
- Combinação dados de diferentes tipos (nominais com racionais, etc);
- Dados com diferentes escalas;
- Dados irrelevantes para a análise;
- Atributos que não contribuem para o aprendizado;
- Dados desbalanceados;
- Número elevado de atributos (maldição da dimensionalidade);
- ...

Como tratar esses problemas?

Técnicas para Tratamento e Transformação de Dados

Técnicas para Tratamento e Transformação de Dados

Tratamento e Transformação de Dados

- **Benefícios**:
  - Facilitar o posterior uso de técnicas de Aprendizado de Máquina.
    - Ex. Alguns métodos consideram somente entradas numéricas e dados normalizados.
  - Permitir comparar atributos de forma adequada com o ajuste da escala dos dados.
  - Redução de complexidade computacional.
    - Ganhos em termos de tempo e custo.
  - Tornar mais fáceis e rápidos ajustes de parâmetros.
  - Facilitar a interpretação dos padrões extraídos.

Tratamento e Transformação de Dados

- **Tratamento e Transformação de Dados**

  - Eliminação manual de atributos,
  - Integração de dados,
  - Amostragem de dados,
  - Redução de dimensionalidade,
  - Balanceamento de dados,
  - Limpeza de dados,
  - Transformação de dados.

Tratamento e Transformação de Dados

- **Eliminação manual de atributos**
  - Alguns atributos não possuem relação com o problema sendo solucionado.
  - Ex.
    - Nome em diagnóstico.
    - Salário entre crianças.

Tratamento e Transformação de Dados

Eliminação manual de atributos:

Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Baixa	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Média	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Não contribuem para estimar se um paciente tem doença ou não

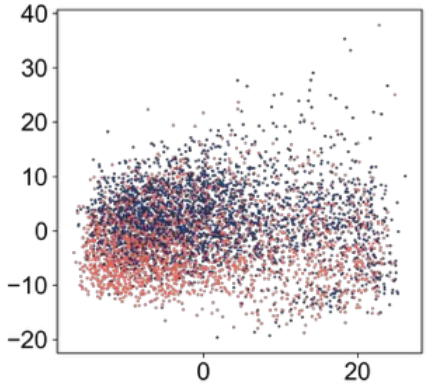
Tratamento e Transformação de Dados

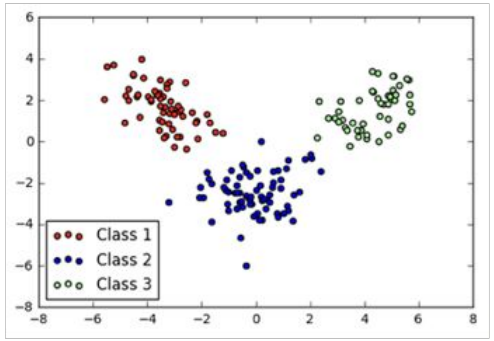
Identificação	Nome	Estado	Idade	Sexo	Nível de Glicose	Febre	Dores no corpo	Diagnóstico
100	João	SP	22	M	90,5	Baixa	Não	Saudável
102	Maria	SP	39	F	86,5	Baixa	Não	Saudável
230	Rubens	RJ	23	M	120,2	Baixa	Sim	Doente
40	João	RS	21	M	85,9	Alta	Não	Doente
543	Marta	SP	40	F	88,7	Alta	Sim	Doente
12	Ana	AM	32	F	78,8	Média	Não	Saudável
130	Carlos	SP	29	M	110	Média	Sim	Doente
123	José	RJ	30	M	80,1	Baixa	Sim	Saudável

Médico pode decidir que atributo associado ao estado de origem do paciente também não é relevante para seu diagnóstico clínico

Tratamento e Transformação de Dados

- **Eliminação manual de atributos**
  - Outro atributo irrelevante facilmente detectado:
    - Atributo que possui o mesmo valor para todos objetos.
  - Há ainda atributos irrelevantes cuja identificação pode ser feita através de técnicas de seleção de atributos.





Tratamento e Transformação de Dados

- **Transformação de valores**
  - Vários algoritmos de AM têm dificuldades em usar os dados em seu formato original.
    - Ex. transformação de valores simbólicos para numéricos.

Primeiro : 1
Segundo: 2
Terceiro: 3

Primeiro	Segundo	Terceiro
1	0	0
0	1	0
0	0	1

One-hot Encoding Scheme

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

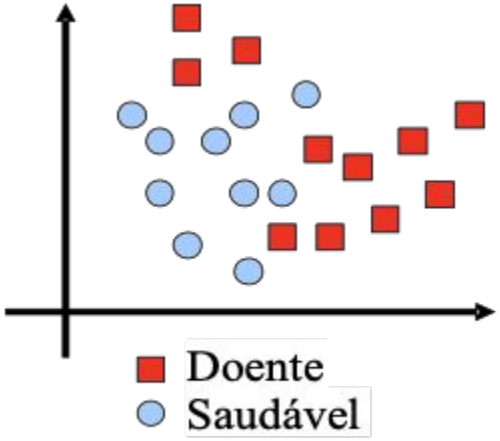
Tratamento e Transformação de Dados

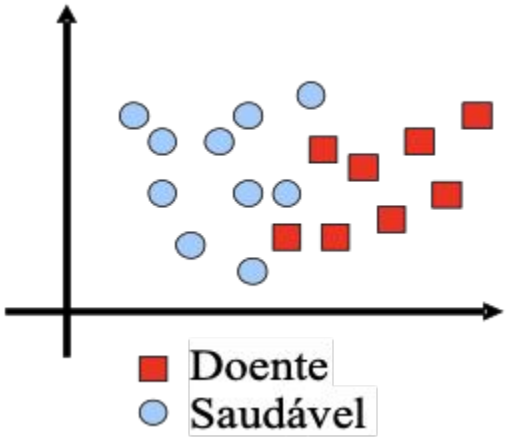
- **Amostragem**
  - Algoritmos de AM podem ter dificuldades em lidar com um grande volume de dados.
    - Saturação de memória.
    - Aumento do tempo computacional para ajustar os parâmetros do modelo.
  - Contudo, quanto mais dados, maior tende a ser a acurácia do modelo.

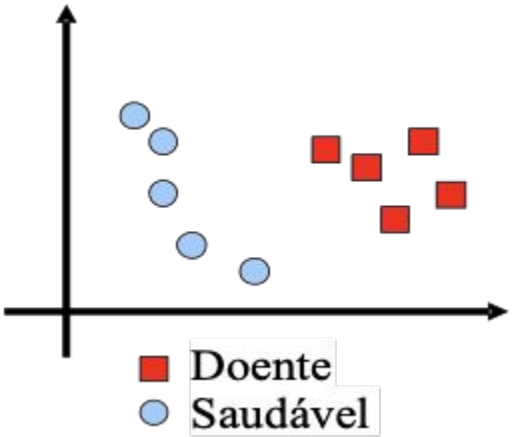
Procurar um balanço entre eficiência computacional e acurácia do modelo.

Tratamento e Transformação de Dados

Amostragem







Tratamento e Transformação de Dados

Amostragem

- Amostragem aleatória simples
  - Variações: com e sem reposição de exemplos (semelhantes quando tamanho da amostra é bem menor que o do conjunto original).
- Amostragem estratificada
  - Quando classes têm propriedades diferentes (ex. números de objetos diferentes).
  - Variações: manter o mesmo número de objetos para cada classe ou manter o número proporcional ao original.
- Amostragem progressiva
  - Começa com amostra pequena e vai aumentando enquanto acurácia preditiva continuar a melhorar.

Tratamento e Transformação de Dados

Amostragem: Balanceamento

- Em muitas aplicações de AM, o número de objetos varia para as diferentes classes.
  - Ex. 80% dos pacientes que vão a um hospital estão doentes
  - Problema na geração/coleta dos dados.
- Vários algoritmos de AM têm o desempenho prejudicado para dados muito desbalanceados, pois tendem a favorecer a classificação na classe majoritária.
- Solução: balancear os dados.

Tratamento e Transformação de Dados

Técnicas de balanceamento

- Acréscimo/eliminação de exemplos na classe minoritária/majoritária
  - Acréscimo: risco de objetos que não representam situações reais e overfitting
  - Eliminação: risco de perda de objetos importantes e underfitting.
- Usar custos de classificação diferentes para as classes.
  - Dificuldades: definição dos custos, incorporar custos em alguns algoritmos de AM.
  - Pode apresentar baixo desempenho quando muitos objetos da classe majoritária são semelhantes.

Tratamento e Transformação de Dados

Limpeza dos dados

- Exemplos de problemas:
  - Ruídos: erros ou valores diferentes do esperado.
    - Ex: Idade com valores negativos.
  - Inconsistências: não combinam/contradizem valores de outros atributos no mesmo objeto.
    - Ex: pessoa com 2m pesando 10 Kg.
  - Redundâncias:
    - Ex: objetos/atributos com mesmos valores.
  - Dados incompletos: ausência de valores de atributos.
    - Ex: Atributo salário para crianças.

Tratamento e Transformação de Dados

- **Dados incompletos: Soluções**
  - Eliminar os objetos com valores ausentes.
  - Definir e preencher manualmente os valores ausentes.
  - Utilizar método/heurística para definir valores automaticamente.
  - Empregar algoritmos de AM que lidam internamente com valores ausentes.

Tratamento e Transformação de Dados

- **Atributos redundantes**
  - Valores podem ser estimados a partir de pelo menos um dos demais atributos.
  - Atributos com a mesma informação preditiva.
    - Ex. atributos idade e data de nascimento
    - Ex. atributos quantidade de vendas, valor por venda e venda total
  - Atributo redundante pode supervalorizar um dado aspecto dos dados.
  - Pode também tornar mais lento o processo de indução.

Transformação de atributos

Transformação de atributos

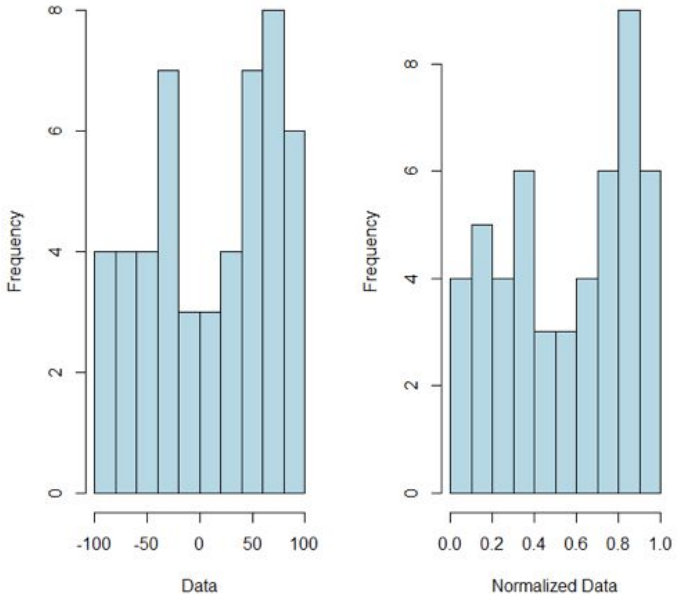
- Algumas vezes é necessário transformar o valor de um atributo numérico em outro valor numérico.
  - Por exemplo, quando o intervalo de valores são muito diferentes, levando a grande variação.
  - Quando vários atributos estão em escalas diferentes.

- As transformações mais comuns são:
  - Normalização (min-max scaling),
  - Padronização,
  - Softmax.

Transformação de atributos

- **Normalização (MinMax Scaling):**
  - Os dados serão ajustados de forma que o valor máximo seja igual a um e o menor, a zero.
  - Usado em redes neurais.
  - Desvantagem: sensível a outliers.

$$Z_i = \frac{X_i - min(X)}{max(X) - min(X)}$$



Transformação de atributos

Normalização:

Idade	Gênero	Peso	Temperatura	Diagnóstico
32	F	79	38	Doente
18	M	95	36,5	Saudável
34	M	92	37	Doente
18	M	87	39	Doente
21	F	45	38,5	Doente
23	M	76	37	Saudável
19	F	56	36,5	Saudável
39	M	109	38	Saudável

Transformação de atributos

Normalização:

Idade	Gênero	Peso	Temperatura	Diagnóstico
0,67	F	0,53	38	Doente
0,00	M	0,78	36,5	Saudável
0,76	M	0,73	37	Doente
0,00	M	0,66	39	Doente
0,14	F	0,00	38,5	Doente
0,24	M	0,48	37	Saudável
0,05	F	0,17	36,5	Saudável
1,00	M	1,00	38	Saudável

$$Z_i = \frac{X_i - 18}{39 - 18} \qquad Z_i = \frac{X_i - 45}{109 - 45}$$

Transformação de atributos

Padronização (Z-score Normalization):

- Os dados ajustados apresentam média igual a zero e desvio padrão igual a um.

$$Z_i = \frac{X_i - \mu_X}{\sigma_X}$$

$$\mu_x = \frac{\sum_{i=1}^{N} X_i}{N}$$

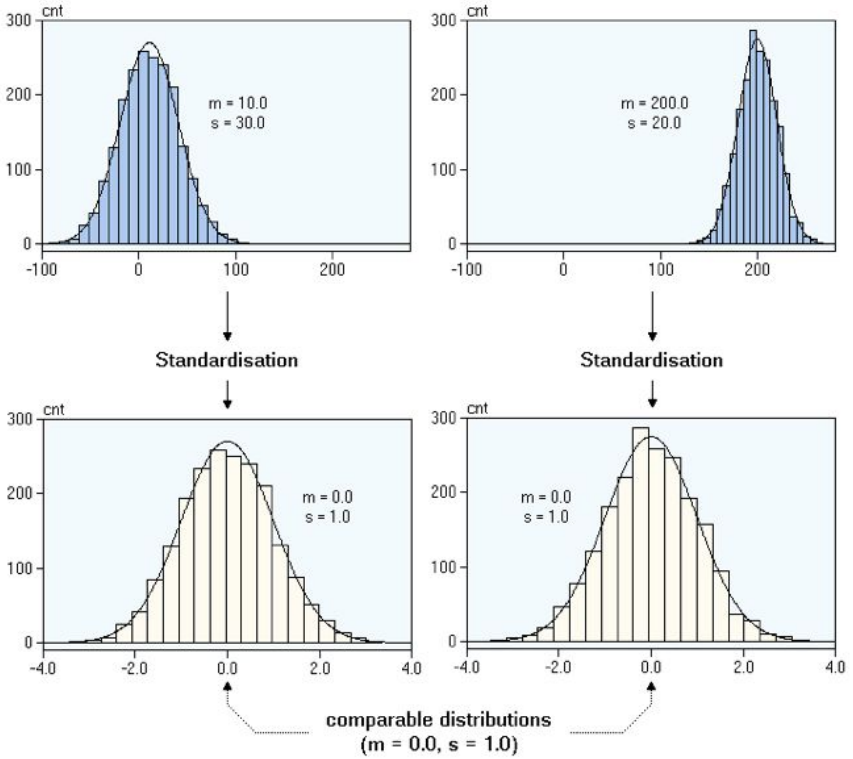
$$\sigma_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2}$$

Transformação de atributos

Padronização

(Z-score Normalization):

$$Z_i = \frac{X_i - \mu_X}{\sigma_X}$$



Transformação de atributos

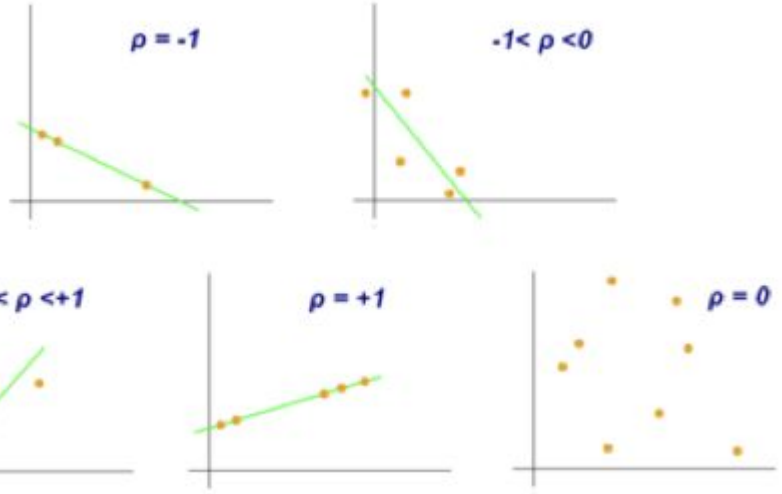
- **Outro tipo de transformação: tradução**
  - Valor é traduzido por um mais facilmente manipulável
    - Ex. converter data de nascimento para idade
    - Ex. converter temperatura de F para C
    - Ex. localização por GPS para código postal.

Análise dos Componentes Principais

Análise dos Componentes Principais

Coeficiente de correlação de Pearson:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$



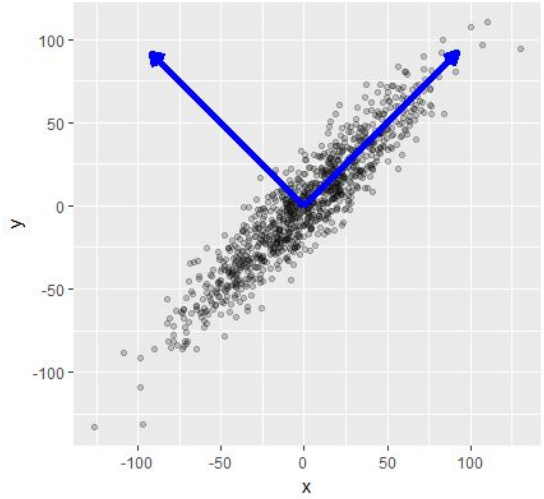
Análise dos Componentes Principais

Motivação: PCA

- Muitos atributos podem ser correlacionados, o que não contribui para a discriminação.
- Um número elevado de atributos pode levar à maldição da dimensionalidade.
- A simplificação dos dados, sem perder informações importantes, ajuda no processamento, pois reduz o tempo computacional e complexidade de algoritmos.
- PCA permite transformar os dados de modo a eliminar redundâncias e preservar informações importantes.

Análise dos Componentes Principais

A análise dos componentes principais é um procedimento para reduzir a dimensão do espaço de variáveis através da transformação dos dados de modo a obter um conjunto de eixos ortogonais (não correlacionados) que capturam grande parte da variabilidade original dos dados.



Análise dos Componentes Principais

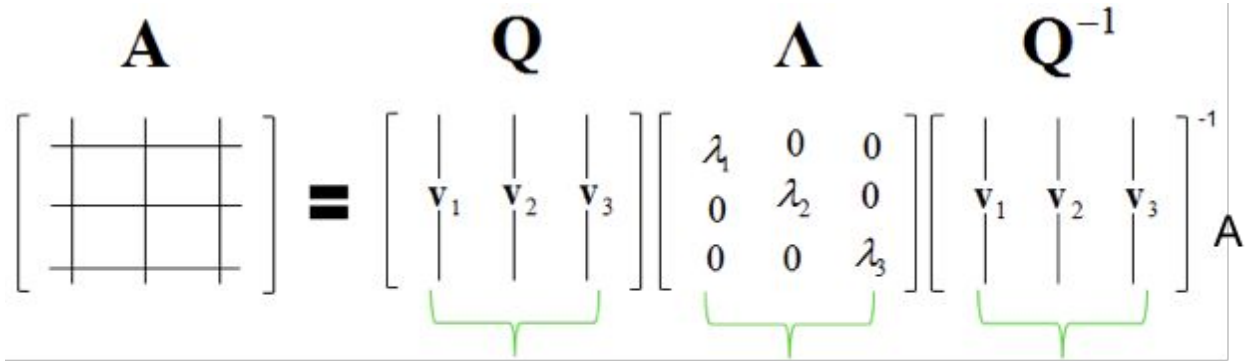
- A matriz de covariância é dada por:

$$\mathbf{A} = cov(X, Y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Para diagonalizarmos uma matriz (encontrar os autovalores e autovetores), usamos o método chamado decomposição em valores singulares (SVD decomposition).

Análise dos Componentes Principais

Decomposição em valores singulares (SVD decomposition):



Autovetores de A | Autovalores de A | Autovetores de A

A inversa Q^{-1} existe apenas se os autovetores são linearmente independentes.

Análise dos Componentes Principais

Passos

1. Centralizar os dados: $Z_i = X_i - \mu_X$
2. Calcular a matriz de covariância.
3. Calcular os autovalores e autovetores da matriz de covariância.
4. Ordenar os autovetores de acordo com o valor dos autovalores.
5. Obter os componentes: multiplicar os dados originais pelos principais autovetores.

Análise dos Componentes Principais

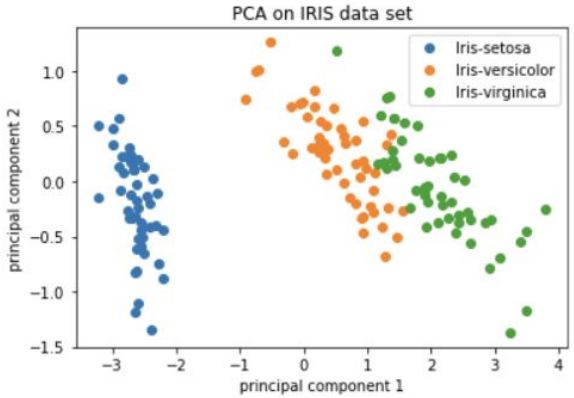
- **Iris dataset:**

	Sepal length	Sepal width	Petal length	Petal width	Class
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	virginica

Iris Versicolor


Iris Setosa


Iris Virginica



Sumário

- Técnicas para Tratamento e Transformação de Dados.
- Normalização
- Análise dos componentes principais.

Leitura Complementar

- Capítulos 2 e 3:

  Inteligência Artificial: Uma abordagem por aprendizado de máquina, Facelli, Lorena, Gamma e Carvalho, LTC.

- PCA:

  A Tutorial on Principal Component Analysis,
  Jonathon Shlens, 2014
  https://arxiv.org/abs/1404.1100