Análise de Componentes Principais

Stefane Adna dos Santos

UFC

Sumário

1. Redução de dimensionalidade
2. Introdução ao PCA
3. Benefícios
4. Cálculo do PCA
5. Prática de PCA

UFC

Redução de dimensionalidade

- Muitos problemas de aprendizagem apresentam uma grande quantidade de atributos.
  - Reconhecimento de imagens;
  - Classificação de textos;
  - Identificação de padrões em dados clínicos/biológicos

UFC

Redução de dimensionalidade

- **Problemas:** Muitos atributos podem resultar em dificuldades na aprendizagem:
  - Dificuldade nos algoritmos de otimização (mais grave);
  - Overfitting (mais grave);
  - Aumento do custo computacional;
  - Maior custo de armazenamento

UFC

Redução de dimensionalidade

- **Solução:**
  - Selecionar ou combinar atributos.

UFC

Redução de dimensionalidade

- **Seleção de atributos:**
  - Visa encontrar um subconjunto de atributos que melhore o desempenho de um algoritmo.

UFC

Redução de dimensionalidade

- **Abordagem independente do modelo (filter):**
  - Define um critério e o usa para selecionar bons atributos a partir dos dados.

- **Abordagem associada a um modelo (wrapper):**
  - Escolhe bons subconjuntos de atributos a partir de seu resultado em um modelo específico.

UFC

Redução de dimensionalidade

- **Abordagem embutida em um modelo (embedded):**
  - O algoritmo de aprendizagem do modelo realiza o treinamento e a seleção de atributos relevantes simultaneamente.

UFC

Redução de dimensionalidade

- **Combinação de atributos:**
  - **Problema:** Selecionar subconjuntos de atributos pode resultar em perda de informação dos dados.
  - **Ideia:** Combinar os atributos originais em vetores de menor dimensão.

UFC

Introdução

- A análise de componentes principais (PCA) é uma técnica estatística utilizada para reduzir a dimensionalidade de conjuntos de dados complexos.

- O objetivo do PCA é encontrar padrões e estruturas subjacentes nos dados, identificando as variáveis que mais contribuem para a variação dos dados.

UFC

Introdução

- Ela busca transformar um conjunto de variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas, chamadas de componentes principais.

UFC

Como funciona?

- Os componentes principais são combinações lineares das variáveis originais ordenadas de forma que a primeira componente principal explique a maior quantidade possível de variação nos dados, a segunda componente principal explique a maior quantidade restante de variação não explicada pela primeira componente, e assim por diante.

UFC

Como funciona?

- O PCA busca resumir as informações contidas em um grande número de variáveis em um número menor de componentes principais, preservando o máximo de variação possível.

UFC

Introdução

- O PCA é amplamente utilizada em diversas áreas, como estatística, análise de dados e aprendizado de máquina.

- Ela pode ser aplicada em diversos tipos de dados, como dados numéricos, dados categóricos e até mesmo em matrizes de covariância ou correlação.

UFC

Introdução

- Ao realizar a análise de componentes principais, é possível realizar diversas tarefas, como redução de dimensionalidade, visualização de dados, detecção de outliers, agrupamento de dados e até mesmo na construção de modelos preditivos.

- Ela fornece uma maneira eficiente de lidar com dados complexos, facilitando a interpretação e a análise dos mesmos.

UFC

Benefícios

- **Redução da dimensionalidade:** O PCA é frequentemente usado para reduzir a dimensionalidade de conjuntos de dados com muitas variáveis. Ele identifica as principais componentes que explicam a maior parte da variação nos dados e projeta os dados em um espaço de menor dimensão, mantendo a maior parte da informação relevante.

UFC

Benefícios

- **Eliminação de multicolinearidade:** Em conjuntos de dados com variáveis altamente correlacionadas, o PCA pode ser usado para identificar as componentes principais não correlacionadas. Isso ajuda a eliminar a multicolinearidade.

UFC

Benefícios

- **Visualização de dados:** O PCA pode ser usado para visualizar dados multidimensionais em um espaço bidimensional ou tridimensional. Ao projetar os dados em um espaço com menos dimensões, é possível plotar os pontos e explorar a estrutura dos dados de forma mais fácil.

UFC

Benefícios

- **Detecção de outliers:** O PCA pode ser usado para identificar outliers nos dados. Os outliers são pontos que se desviam significativamente do padrão geral dos dados e podem ser identificados através da análise das distâncias entre os pontos projetados nas principais componentes.

UFC

Benefícios

- **Compressão de dados:** O PCA também pode ser usado para comprimir dados, reduzindo a quantidade de informações necessárias para representar os dados originais. Ao projetar os dados nas principais componentes, é possível armazenar apenas as informações mais importantes e reconstruir os dados originais com uma perda mínima de informação.

UFC

Benefícios

- **Pré-processamento de dados:** O PCA pode ser usado como uma etapa de pré-processamento antes da aplicação de outros algoritmos de aprendizado de máquina. Ele pode ajudar a melhorar o desempenho e a eficiência de outros métodos, removendo o ruído e a redundância dos dados.

UFC

Cálculo do PCA

- **Passo 1 - Normalização dos dados:**
  - É importante normalizar os dados para garantir que todas as variáveis tenham a mesma escala. Isso pode ser feito através do cálculo da média e do desvio padrão de cada variável e aplicando uma transformação z (normalização z-score).

UFC

Cálculo do PCA

- **Passo 2 - Cálculo da matriz de covariância ou matriz de correlação:**
  - Com os dados normalizados, podemos calcular a matriz de covariância ou matriz de correlação, dependendo do contexto.

UFC

Cálculo do PCA

- **Passo 3 - Decomposição da matriz:**
  - Realizamos a decomposição da matriz de covariância ou matriz de correlação para obter os autovetores e autovalores correspondentes.

UFC

Cálculo do PCA

- **Passo 4 - Ordenação dos componentes principais:**
  - Ordenamos os autovetores de acordo com os autovalores correspondentes, de forma decrescente. Os autovetores com os maiores autovalores explicam a maior parte da variação nos dados.

UFC

Cálculo do PCA

- **Passo 5 - Projeção dos dados:**
  - Projetamos os dados originais nos componentes principais selecionados, obtendo assim os dados transformados em um espaço de menor dimensionalidade.

UFC

Cálculo do PCA

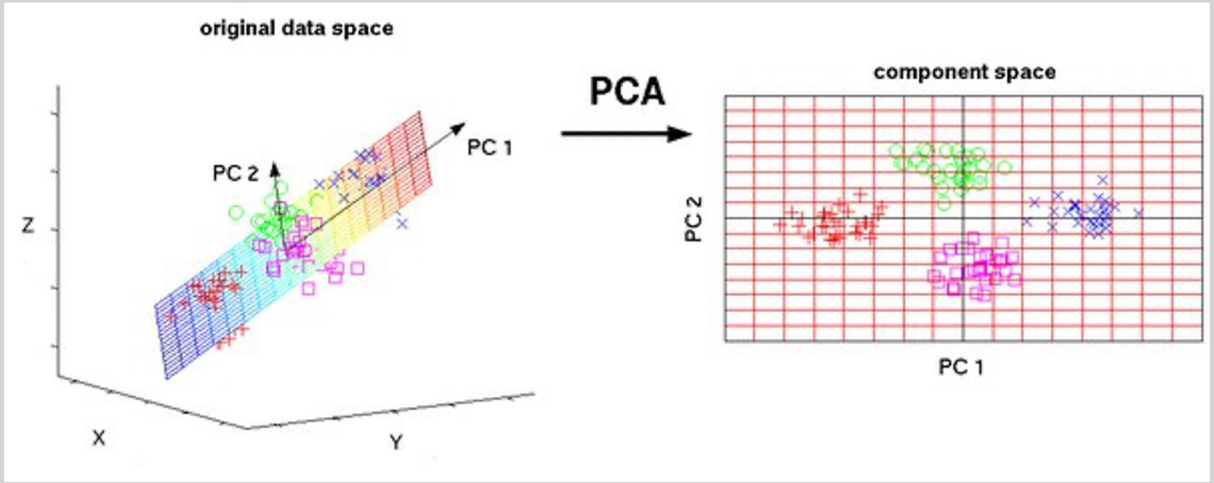

Figura 01: PCA

UFC

Prática

- Sklearn:
  - from sklearn.decomposition import PCA

- Código:
  - https://github.com/stefaneadna/estagio_a_docencia_ciencias_de_dados/tree/main/PCA

UFC

Avaliação da disciplina

- https://forms.gle/1KGrkPncso54tkUu7

UFC