PUC-SP

Mineração de Dados

Prof. Dr. Daniel Rodrigues da Silva

Mineração de Dados

PCA

PUC-SP

Bibliografia básica:

Introdução à mineração de dados : conceitos básicos, algoritmos e aplicações. Leandro Nunes de Castro e Daniel Gomes Ferrari. – São Paulo : Saraiva, 2016.

Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina . André Carlos Ponce de Leon Ferreira et al. 2 Ed. LTC, 2024.

Principal Component Analysis (PCA): O PCA consiste em uma técnica estatística que utiliza uma transformação linear para re-expressar um conjunto de atributos em um conjunto menor de atributos linearmente independentes, mantendo boa parte das informações contidas nos dados originais [Dunteman 1989].

Em suma, os objetivos do PCA consistem em [Abdi e Williams 2010]:

- Extrair apenas as informações mais relevantes de um conjunto de dados;
- Comprimir o tamanho do conjunto de dados original, simplificando sua descrição;
- Permitir a análise da estrutura dos objetos e dos atributos de um conjunto de dados.

Em outras palavras:

PCA, reduz o número de dimensões em grandes datasets aos componentes principais mantendo a maior parte das informações originais.

Ela faz isso transformando variáveis potencialmente correlacionadas em um conjunto menor de variáveis, chamadas **componentes principais.**

PUC-SP

PCA é eficaz na visualização e exploração de datasets de alta dimensão ou conjuntos com muitos atributos, porque identifica tendências, padrões e/ou valores discrepantes.

PCA extrai os atributos mais importantes de grandes conjuntos de dados, enquanto preserva as informações mais relevantes do conjunto de dados original.

Isso diminui a complexidade do modelo, a medida que a adição de cada novo atributo impacta negativamente o desempenho do modelo, o que é geralmente chamado de "***maldição da dimensionalidade***".

Quando projeta um conjunto de dados de alta dimensão em um espaço com menos atributos, o PCA também minimiza ou elimina completamente problemas comuns, tais como:

- ***Multicolinearidade***

A multicolinearidade acontece quando duas ou mais variáveis independentes estão altamente correlacionadas entre si, o que pode ser um problema para a modelagem.

PUC-SP

- *Sobreajuste (overfitting)*

Modelos sobreajustados tendem a generalizar os novos dados de modo não satisfatório, diminuindo potencialmente seu valor.

PUC-SP

Existem variações do método PCA, tais como:

- ***Regressão por componentes principais e***
- ***Kernel PCA.***

Aqui, porém, nos concentraremos no método tradicional encontrado na literatura atual.

PUC-SP

PCA reduz a dimensionalidade em datasets, como também a denominada *"Análise Discriminante Linear (LDA)"*.

Porém, ao contrário da LDA, a PCA não se limita a tarefas de Aprendizado Supervisionado, mas também é útil no Aprendizado não Supervisionado, pois consegue reduzir as dimensões sem necessidade da existência de rótulos de classe ou categorias.

PCA reduz as variáveis a um subconjunto de componentes principais linearmente independentes.

PUC-SP

Existem ainda outras técnicas de redução de dimensionalidade, tais como:

- *Análise Discriminante Linear (LDA)*
- *A floresta aleatória*
- *A aproximação e projeção múltipla uniforme (UMAP) e*
- *Incorporação de vizinhos estocásticos distribuídos em t (t-SNE).*