Principal Component Analysis - PCA

Gustavo Aparecido de Souza Viana

***Abstract*— A área de machine learning vem crescendo exponencialmente, com intuíto de desenvolver algoritmos ou aplicações para analisar grande bases de dados e assim tirar conclusões, no entanto, podemos encontrar nesse caminho algumas bases que contém muitas caracteristicas e poucas observações de exemplos. Com base nesse contexto, o PCA foi criado para reduzir essa proporção de Caracteristicas X Observações, assim podemos reduzir a dimensão do problema original. Os resultados mostraram que o PCA é capaz de reduzir a dimensão e além disso, consegue explicar ou representar o mesmo ou quase por completo o dado com o número de dimensões reduzido.**

I. INTRODUÇÃO

A necessidade de predições está sendo bem comum e usada na atualidade, podemos dar o exemplo de predições relacionadas ao mercado de ações ou até mesmo predições relacionadas a área médica, prever se o paciente tem tendências de desenvolver algum tipo de câncer.

Conseguimos predizer algum informação baseada em uma base conhecida com machine learning ou com métodos estatísticos. É de extrema importância conhecer o ramo de atividade da aplicação pois métodos de machine learning geralmente estão relacionados a rede neural, consquentemente necessitam de mais dados para que as predições tenham uma maior acurácia. De contra partida, métodos estatísticos não necessitam.

Neste trabalho foi abordado a implementação do *Principal Component Analysis* com objetivo de reduzir o número de dimensões de uma base de dados para assim conseguirmos representar majoritariamente o dado. Para testar essas técnicas foram utilizados três bases *Alps Water*, *Books x Grades* e *US Census Dataset*.

II. CONCEITOS FUNDAMENTAIS

Nesta seção, será apresentado os conceitos básicos utilizados para o *Principal Component Analysis* que incluí a Covariância e Autovalores e Autovetores.

A. Covariância

A covariância na estatítisca, ou variância conjunta, é uma medida do grau de interdependência numérica entre duas variáveis aleatórias [2]. Ela pode ser positiva e negativa ou zero, positiva quando a variação tem o mesmo sentido, negativa quando são sentidos opostos e zero quando não há variação. A Equação 1 representa o cálculo entre x e y, onde e representa a média da sua própria população, e N a quantidade de observações.

$$\sqrt{\sum \frac{(x_i - \hat{x_i})(y_i - \hat{y_i})}{N-1}} \tag{1}$$

B. Autovalor e Autovetor

Autovetor e autovalor, são vetores e valores respectivamente de uma matriz onde suas multiplicações não alteram a direção, apenas a magnitude [3]. Sendo assim, dado um λ, sendo um valor escalar, multiplicado por um vetor X de uma matriz quadrada A, podemos dizer que λ é autovalor da matriz A caso exista um vetor $! = 0$ tal que $Ax = \lambda x$, sendo assim o vetor x é chamado de autovetor. Abaixo a Equação 2 representa o conceito do autovetor e valor, e a Equação **??** representa um exemplo de autovalor e autovetor.

$$Matriz_a Vetor_i = \lambda Vetor_i \tag{2}$$

$$\begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} \begin{vmatrix} 3 \\ 2 \end{vmatrix} = 4 \begin{vmatrix} 3 \\ 2 \end{vmatrix} \tag{3}$$

C. Principal Component Analysis

A ideia do *PCA* é reduzir a dimensionalidade dos dados para assim representar o mesmo dado com menos dimensões que são novas variáveis linearmente independentes chamadas de componentes [1].

O processo de conversão inicia pelo cáculo da covariância, e após isso conseguimos calcular os autovetores e autovalores. Os autovalores demonstram o quanto aquela componente representa ou explica do dado em si, quanto maior o autovalor mais importante e consequentemente, explica majotariamente o dado. O autovetor é a componente em si, uma matriz de transformação, ou seja, se multiplicarmos os dados pelo autovetor teremos uma nova base reduzida. Imporante ressaltar que cada componente é ortogonal a anterior e assim sucessivamente.

III. METODOLOGIA

Nesta seção será apresentado a metodologia utilizada para implementar o *Principal Analysis Component*, implementados em Python. O código fonte pode ser encontrado em https://github.com/apparecidoo/master-special-learning-topic.

A metodologia é consistida em 7 passos. Iniciando com a leitura da base de dados citada na introdução, em seguida é feita para cada observação a subtração da média delas. No terceiro passo é calculado a variância da matriz resultante. No quarto e quinto passo é calculado o auto valor e auto vetor, respectivamente. No sexto passo são selecionadas as componentes que maior representam o dado e por fim no último passo as componentes são aplicadas na base original à fim de transforma-la.

IV. EXPERIMENTOS E RESULTADOS

Nesta seção será apresentado os experimentos e seus respectivos resultados.

O experimento consiste na aplicação do *Principal Component Analysis* para cada base de dados, sendo elas a *Alps Water*, *Books x Grades* e *US Census*.

As Figuras 1, 2 e 3 representam o quanto cada componente representa para cada base de dados.

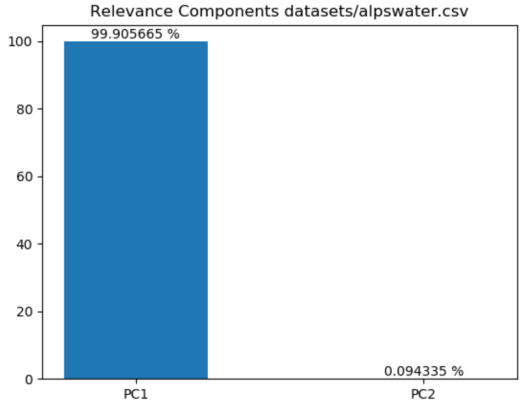

Fig. 1: Representação de cada componente para o dataset *Alps Water*

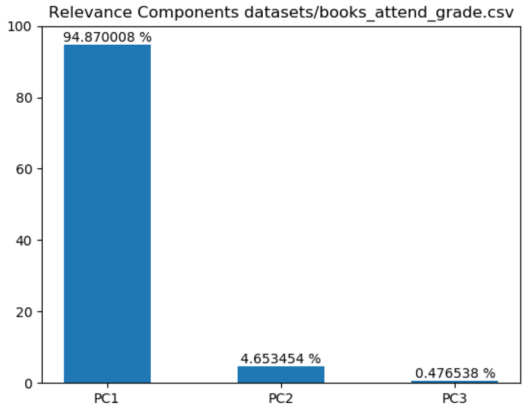

Fig. 2: Representação de cada componente para o dataset *Books x Grades*

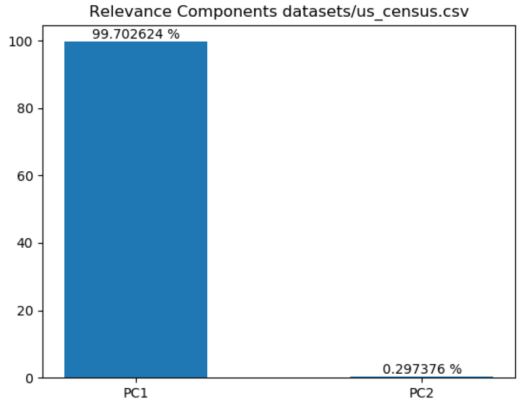

Fig. 3: Representação de cada componente para o dataset *Us Census*

As Figuras 4, 5 e 6 representam os dados orignais e transformados da base de dados original plotados.

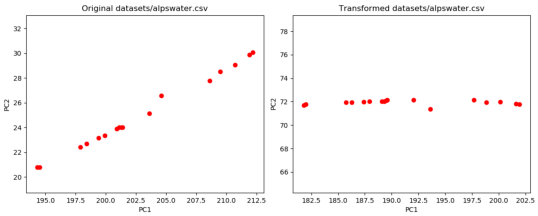

Fig. 4: Representação dos dados originais e dados transformados do dataset *Books x Grades*

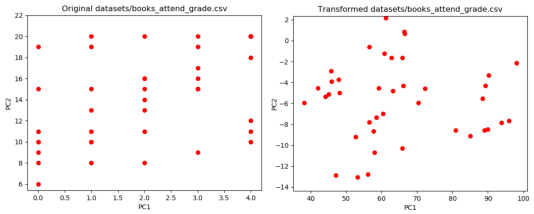

Fig. 5: Representação dos dados originais e transformados do dataset *Alps Water*

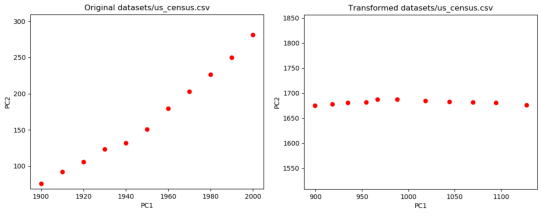

Fig. 6: Representação dos dados originais e transformados do dataset *US Census*

V. CONCLUSÃO

Após a análise do conceito e resultados obtidos após a aplicação dos métodos e comparando os mesmos, podemos concluir que o *PCA* consegue reduzir a dimensionalidade,

e além disso podemos representar o mesmo dado da base somente com **UMA** única componente, e que estamos falando de base de dados com uma quantidade pequena de observações e caracteristícas, ou seja, em termos de processamento isso não muda tanto mas quando temos uma base de dados muito ampla, a diferença é muito significante.

REFERENCES

[1] M. Baxter and M. Heyworth. Principal components analysis of compositional data in archaeology. 1989.

[2] P. J. Bickel and E. Levina. Covariance regularization by thresholding. 2008.

[3] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. 1985.