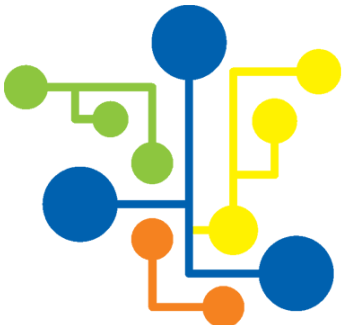


Data Science Academy

www.datascienceacademy.com.br

Matemática Para Machine Learning

E-book
Eigenvectors, Eigenvalues, PCA, Covariance e Entropy
Parte 2

Na Parte 2 deste e-book vamos compreender o que é o PCA (Principal Component Analysis ou Análise dos Componentes Principais).

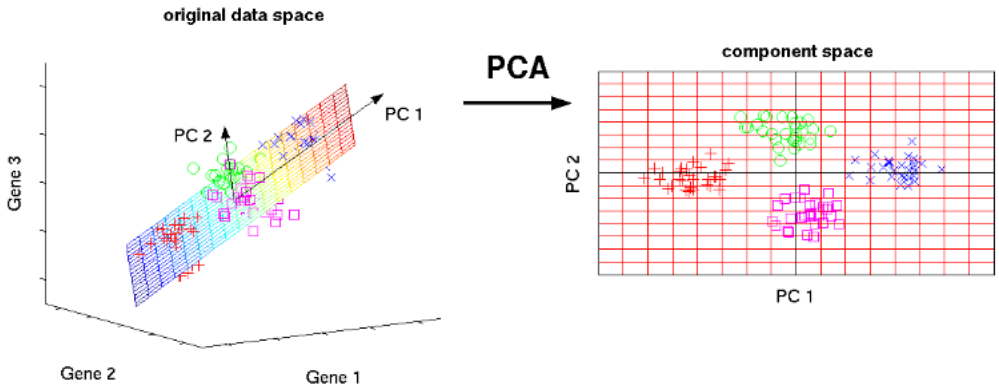
O PCA é uma ferramenta para encontrar padrões em dados de alta dimensão, como imagens. Profissionais de aprendizado de máquina às vezes usam o PCA para pré-processar dados para suas redes neurais. Ao centralizar, girar e dimensionar dados, o PCA prioriza a dimensionalidade (permitindo que você elimine algumas dimensões de baixa variação) e pode melhorar a velocidade de convergência da rede neural e a qualidade geral dos resultados.

Para chegar ao PCA, vamos definir rapidamente algumas ideias estatísticas básicas - média, desvio padrão, variância e covariância - para que possamos juntá-las mais tarde. Suas equações estão intimamente relacionadas. Primeiro, vamos definir o PCA.

O Que é o PCA?

O PCA foi inventado em 1901 por Karl Pearson e utiliza álgebra linear para transformar datasets em uma forma comprimida, o que é geralmente conhecido como Redução de Dimensionalidade. Com PCA você pode escolher o número de dimensões (chamados componentes principais) no resultado transformado.

A Análise de Componentes Principais (PCA) é um método para extração das variáveis importantes (na forma de componentes) a partir de um grande conjunto de variáveis, disponíveis em um conjunto de dados. Esta técnica permite extrair um número pequeno de conjuntos dimensionais a partir de um dataset altamente dimensional. Com menos variáveis a visualização também se torna muito mais significativa. PCA é mais útil quando se lida com 3 ou mais dimensões.



Cada componente resultante é uma combinação linear de n atributos. Cada componente principal é uma combinação de atributos presentes no dataset. O Primeiro Componente Principal é a combinação linear dos atributos com máxima variância e determina a direção em que há mais alta variabilidade nos dados. Quanto maior a variabilidade capturada no primeiro componente principal, mais informação será capturada pelo componente. O Segundo Componente Principal captura a variabilidade remanescente. Todos os componentes subsequentes possuem o mesmo conceito.

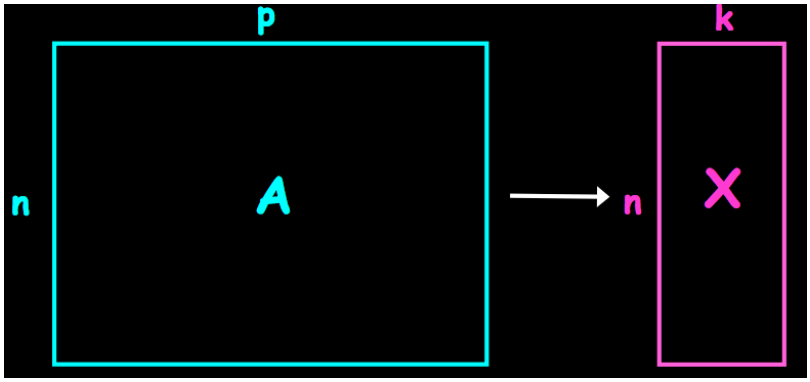
O PCA precisa ser alimentado com dados normalizados. Utilizar o PCA em dados não normalizados pode gerar resultados inesperados.

A análise de componentes principais é uma técnica da estatística multivariada (e que será estudada em um capítulo inteiro no curso Análise Estatística Para Data Science II) que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis denominadas de componentes principais. Os componentes principais apresentam propriedades importantes: cada componente principal é uma combinação linear de todas as variáveis originais, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados. Os componentes principais são garantidamente independentes apenas se os dados forem normalmente distribuídos.

Procura-se redistribuir a variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados. Esta técnica pode ser utilizada para geração de índices e agrupamento de indivíduos. A análise agrupa os indivíduos de acordo com sua variação, isto é, os indivíduos são agrupados segundo suas variâncias, ou seja, segundo seu comportamento dentro da população, representado pela variação do conjunto de características que define o indivíduo, ou seja, a técnica agrupa os indivíduos de uma população segundo a variação de suas características.

A análise de componentes principais é associada à ideia de redução de massa de dados, com menor perda possível da informação.

O objetivo é sumarizar os dados que contém muitas variáveis (p) por um conjunto menor de variáveis (k) compostas derivadas a partir do conjunto original. PCA usa um conjunto de dados representado por uma matriz de n registros por p atributos, que podem estar correlacionados, e sumariza esse conjunto por eixos não correlacionados (componentes principais) que são uma combinação linear das p variáveis originais. As primeiras k componentes contém a maior quantidade de variação dos dados.



Em termos gerais o PCA busca reduzir o número de dimensões de um dataset, projetando os dados em um novo plano. Usando essa nova projeção os dados originais, que podem envolver diversas variáveis, podem ser interpretados utilizando menos "dimensões."

No dataset reduzido podemos observar com mais clareza tendências, padrões e/ou outliers. Mas vale lembrar que a regra: "Se não está nos dados brutos não existe!" é sempre válida. PCA fornece apenas mais clareza aos padrões que já estão lá.

Detalhes Estatísticos e Matemáticas do PCA

A média é simplesmente o valor médio de todos os **xs** no conjunto X, que é encontrado dividindo a soma de todos os pontos de dados pelo número de pontos de dados, n.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

O desvio padrão, por mais divertido que isso pareça, é simplesmente a raiz quadrada da distância quadrada média dos pontos de dados até a média. Na equação abaixo, o numerador contém a soma das diferenças entre cada ponto de dados e a média, e o denominador é simplesmente o número de pontos de dados (menos um), produzindo a distância média.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}$$

A variação é a medida do spread dos dados. Se eu pegar uma equipe de jogadores de basquete holandeses e medir sua altura, essas medições não terão muita variação.

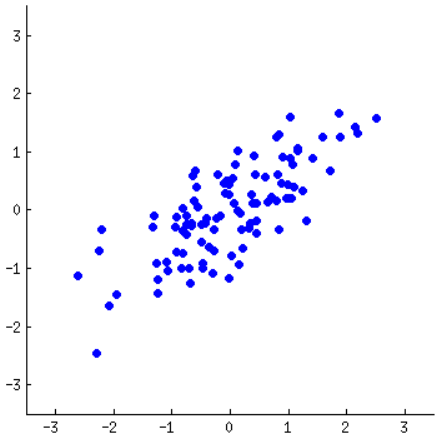
Mas se eu jogar o time de basquete holandês em uma sala de aula de alunos de jardim de infância, então as medidas de altura do grupo combinado terão muita variação. Variância é o spread ou a quantidade de diferença que os dados expressam.

A variação é simplesmente o desvio padrão ao quadrado e é frequentemente expressa como s^2.

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

Tanto para a variação quanto para o desvio padrão, a correspondência entre as diferenças entre os pontos de dados e a média as torna positivas, de modo que os valores acima e abaixo da média não se anulam mutuamente.

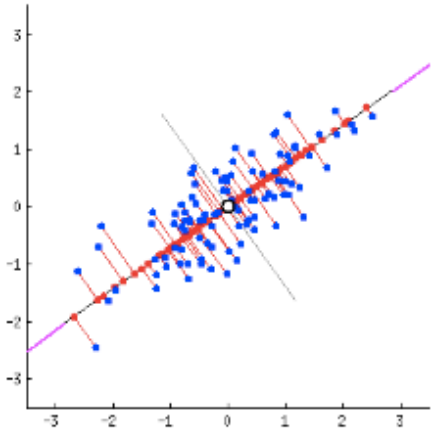
Suponhamos que você tenha plotado a idade (eixo x) e a altura (eixo y) desses indivíduos (definindo a média como zero) e criou um gráfico de dispersão:



O PCA tenta desenhar linhas diretas e explicativas através de dados, como a regressão linear.

Cada linha reta representa um “componente principal” ou uma relação entre uma variável independente e dependente. Embora haja tantos componentes principais quanto dimensões nos dados, o papel do PCA é priorizá-los.

O primeiro componente principal divide um gráfico de dispersão com uma linha reta de uma maneira que explica a maior variação; isto é, segue a dimensão mais longa dos dados. (Isso acontece para coincidir com o menor erro, conforme expresso pelas linhas vermelhas ...). No gráfico abaixo, ele corta o comprimento da reta.



O segundo componente principal corta os dados perpendiculares ao primeiro, ajustando os erros produzidos pelo primeiro. Existem apenas dois componentes principais no gráfico acima, mas se fosse tridimensional, o terceiro componente se encaixaria nos erros do primeiro e do segundo componente principal, e assim por diante.

Consulte as referências abaixo. Continuamos na Parte 3.

Referências:

https://math.stackexchange.com/questions/24456/matrix-multiplication-interpreting-and-understanding-the-process/24469#24469

https://pdfs.semanticscholar.org/9dfa/3d30681788aac5077ede7b0ba2f7c4ac501e.pdf

https://news.ycombinator.com/item?id=10080415

https://skymind.ai/wiki/eigenvector

https://www.cs.cmu.edu/~mgormley/courses/10601-s17/slides/lecture18-pca.pdf

https://arxiv.org/pdf/1407.2904.pdf

http://www.uta.fi/sis/mtt/mtts1-dimensionality_reduction/drv_lecture3_jan28update.pdf

http://mathworld.wolfram.com/MatrixDiagonalization.html