UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Física
Bacharelado em Engenharia Física
Tópicos especiais de engenharia física

Deomar Santos da Silva Junior / 00260682

Segmentação e análise de características de cânceres de mama malignos e benignos por Análise de Componentes Principais (PCA)

Introdução

O câncer de mama é o tipo de câncer mais comum entre as mulheres no mundo e no Brasil, após o câncer de pele não melanoma. Para o ano de 2018 foram estimados 59700 novos casos de câncer de mama no Brasil [1]. O câncer pode ser classificado como maligno, em que o crescimento anormal de células se restringe a uma camada fibrosa, ou maligno, em que o câncer pode entrar na corrente sanguínea e crescer em outras partes do corpo, tipo que causa a maior parte das mortes [2]. Com o objetivo de identificar o tipo do câncer, foi feita análise de componentes principais no dataset de câncer de mama de Wisconsin [3] que agrega 30 diferentes variáveis de 569 tipos de cânceres de mama como, por exemplo, o raio do tumor, média do perímetro, contornos etc.

Análise de Componente Principal (PCA)

A PCA é uma técnica matemática que permite que, a partir de várias variáveis, seja possível filtrar a combinação dessas variáveis que são mais relevantes para o conjunto de dados analisados. Ou, na linguagem da álgebra linear, é uma transformação de um conjunto de observações correlacionadas em um conjunto de variáveis linearmente não correlacionadas [4], as quais são chamadas de componentes principais de variáveis.

Por exemplo, dado um conjunto de dados originais sobre pessoas como, por exemplo, altura (*feature 1*) e a capacidade para entrar em espaços pequenos (*feature 2*). - figura 1. Caso queira se saber, por exemplo, baseado nos dados apresentados, o sexo biológico da pessoa, pode ser pensado que a probabilidade da pessoa ser um homem aumenta quanto maior a altura e menor a capacidade de entrar em espaços pequenos. Como se deseja saber as direções que produzem a

maior variação nos dados, pode-se dividir o conjunto de dados em duas componentes. A componente 1 (*Component 1*) apresenta a maior variação, enquanto a componente 2 (*Componente 2*) apresenta a menor variação. Pode-se entender a componente 1 como uma combinação linear entre a altura e a capacidade de entrar em espaços pequenos de forma que quanto maior a feature 1 e menor a 2, "caminha-se" para a direita na componente 1. Enquanto a componente 2 poderia ser entendida como uma combinação linear entre as duas variáveis que não modifica muito a posição do dado no gráfico.

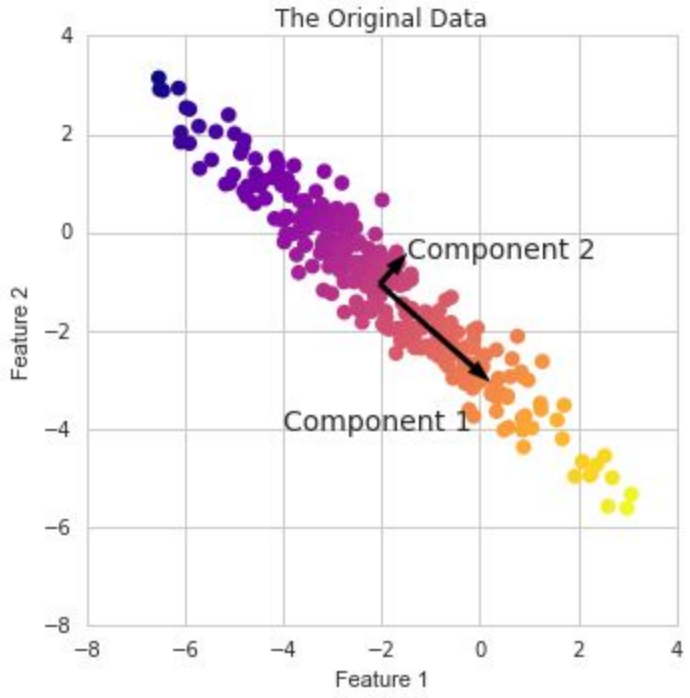

Figura 1. Duas *features* diferentes para um dataset genérico [5].

Desta forma, identificando-se as componentes principais que apresentam a maior variabilidade dos dados, pode-se reduzir a dimensão dos dados filtrando-se apenas a nova direção (figura 2).

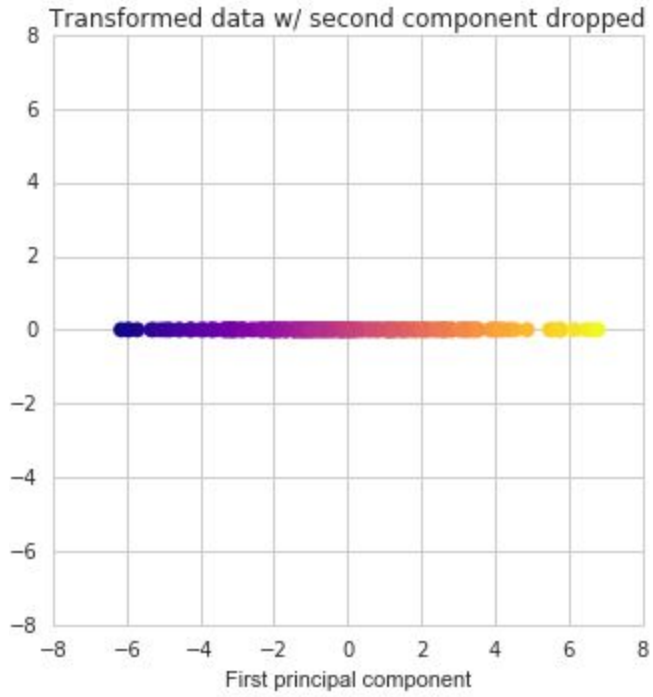

Figura 2. Componente principal filtrada do conjunto de dados [5].

Para encontrar a combinação linear que gera as componentes 1 e 2, calcula-se a correlação matemática entre as duas features, que agora são entendidos como vetores, e então maximiza-se os vetores de correlação. Como resultado se tem os autovetores do conjunto que indicarão as direções das componentes 1 e 2.

Matematicamente, para alcançar os resultados mencionados anteriormente, aplica-se o seguinte procedimento:

Dado uma matriz $X = r_{ij}$ com os dados experimentais onde cada coluna $\boldsymbol{j}$ representa uma *feature* e cada linha $\boldsymbol{i}$ um ponto amostral diferente para cada *feature*. Para avaliar a variação de cada *feature*, deve-se primeiramente normalizar a matriz de *features* uma vez que se quer eliminar os efeitos de magnitude de uma *feature* em relação a outra.

Então:

$$X = (r_{i,j} - < r >_j) / \sqrt{< r^2 >_j - < r^2 >_j^2}$$

Onde $< r >_j$ é a média da *feature* $\boldsymbol{j}$.

Após, calcula-se a matriz de correlação (C_{ij}) entre os dados:

$$C_{ij} = X^T X$$

Então, minimiza-se a matriz de correlação de forma que a matriz de dados seja uma distribuição de média 0. Para essa maximização com condição se utiliza o multiplicador de Lagrange (L):

$$L = X^T CX - \mu(X^T X - 1I)$$

Como resultado se obtém a equação de autoestados:

$$C\alpha = \mu\alpha$$

Ou seja, os autovetores da equação de autoestados gera os vetores de maximização da matriz de correlação.

Por fim, para se obter as saídas correspondentes das componentes (T):

$$T = X\alpha$$

Onde $\propto$ representa os autovetores.

PCA aplicada no dataset de Wisconsin

Aplicando-se as equações acima no dataset do câncer de Wisconsin e plotando-se as duas componentes principais, tem-se:

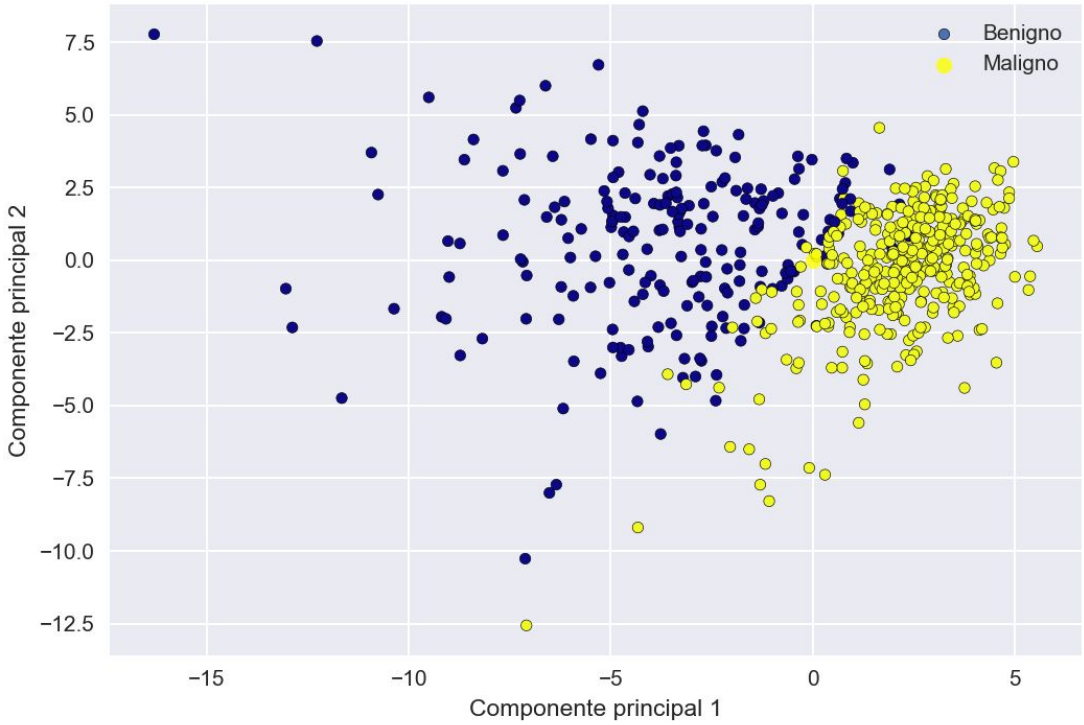

Figura 3. Componentes principais filtradas do conjunto de dados [5].

Observa-se claramente que as duas componentes reproduzem bem a maior variabilidade do conjunto de dados de forma que é possível, apenas a partir de duas features reduzir em 28 dimensões o problema original e ainda assim dividir claramente o conjunto alvo de câncer maligno e benigno. Para efeito de comparação, plota-se (figura 4) as duas variáveis "*radius error*" e "*perimeter error*" com 0.56 e 0.55 de correlação com os dados originais que, apesar da alta correlação, não dividem os dados de forma tão clara como as componentes principais da figura 3.

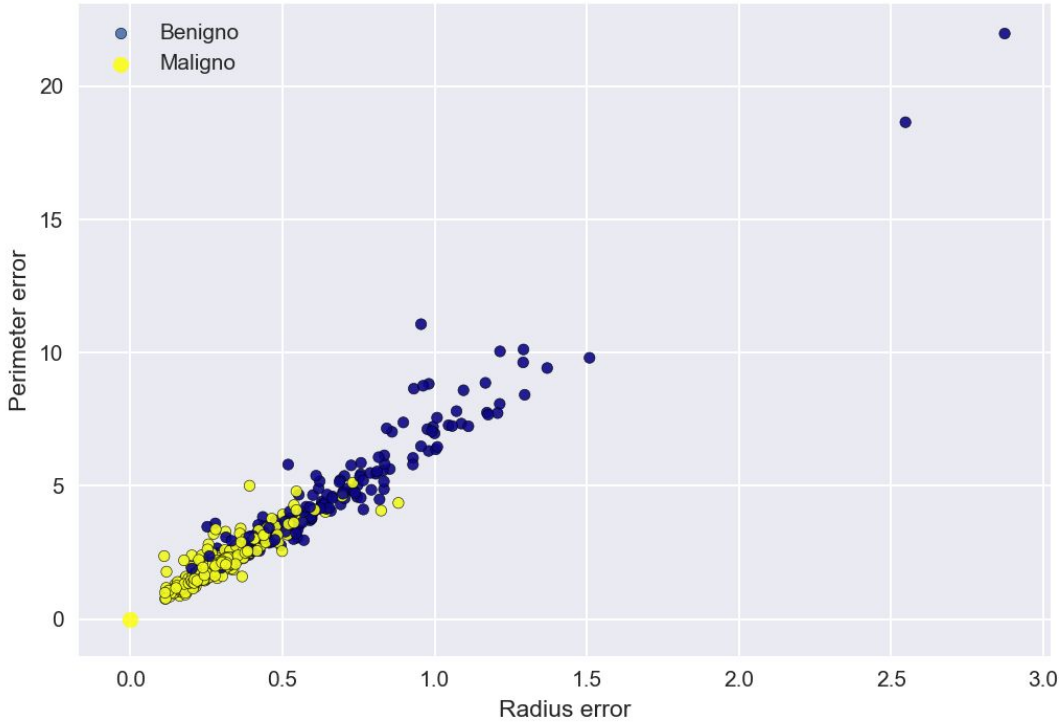

Figura 4. Duas variáveis de alta correlação em relação aos alvos de câncer maligno e benigno.

Variância explicada

A variância explicada é uma medida relativa de quanto a componente principal "explica" os dados. É definida por:

$$\sigma^2_{exp} = \lambda_n / \sum_i^N \lambda_i$$

De forma que a magnitude relativa do autovalor λ_n associado ao autovetor $\propto_n$ em relação a soma de todos os autovalores representa o peso do autovetor ou componente para representar os dados.

Para o conjunto das duas componentes principais expostas acima, tem-se, para a componente 1 e 2, respectivamente:

σ^2_{exp} = 0.44 e 0.19

Explicando, então, 44% e 29% dos dados.

Os respectivos autovalores são 13.28 e 5.69.

Os autovalores associados aos outros 28 autovetores podem ser vistos no programa em python anexado.

Eixos relacionados a ruídos

Como há 30 *features* no dataset, há muitas opções de componentes que representam boa parte dos dados e igualmente componentes que podem ser representadas como ruídos. Então, para exemplificar a diferença, em contraste com as duas componentes principais, as duas componentes com a menor variância explicada são plotadas na figura 5.

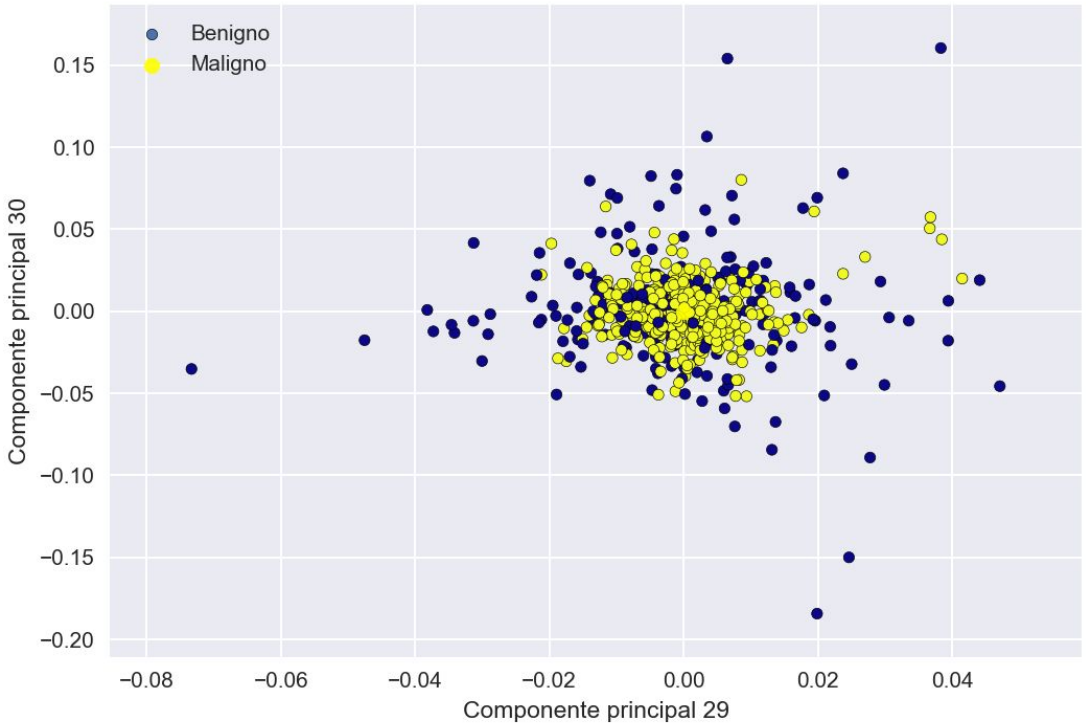

Figura 5. Componentes que representam os ruídos em relação aos alvos de câncer maligno e benigno.

Percebe-se que não é possível traçar uma linha que divida o dataset nos dois tipos de câncer devido à baixa influência dessas duas componentes no dataset. A variância das PCAs 29 e 30 são, respectivamente, 0.00074 e 0.00013.

Kernel análise das principais componentes (KPCA)

O método PCA abordado anteriormente filtra apenas as componentes lineares do conjunto de dados, de forma que se desejar separar as componentes não lineares que produzem a maior variação não será possível. Entretanto o método KPCA permite que sejam filtradas as componentes não lineares.

Matematicamente, em vez de construir a matriz de correlação C_{ij}, aplica-se uma função *kernel* no conjunto de dados X:

Assim, a matriz C vira a matriz K:

$$K = \sum_{i}^{N} f(r_i) f(r_i)^T$$

Aplicando-se os mesmos passos anteriores, tem-se, ao final:

$$K\alpha = \lambda_n \alpha$$

Ou seja, novamente se tem a matriz de autoestados que gera as componentes principais.

É importante salientar que, como no caso anterior, a matriz de correlação é centrada em torno de um valor com média 0, a matriz *kernel* deve obedecer a mesma condição. Então, antes de resolver a equação de autoestados é necessário centralizar a matriz kernel.

Por fim, para gerar a matriz de saída com as componentes principais (T), multiplica-se os autovetores pela matriz kernel:

$$T = \alpha K$$

O kernel utilizado neste trabalho é o *kernel* gaussiano, de forma que pode-se substituir K por:

$$K = exp(-\beta ||r_i - r_j||^2$$

Então calculando-se as duas componentes principais para o *kernel* gaussiano tem-se:

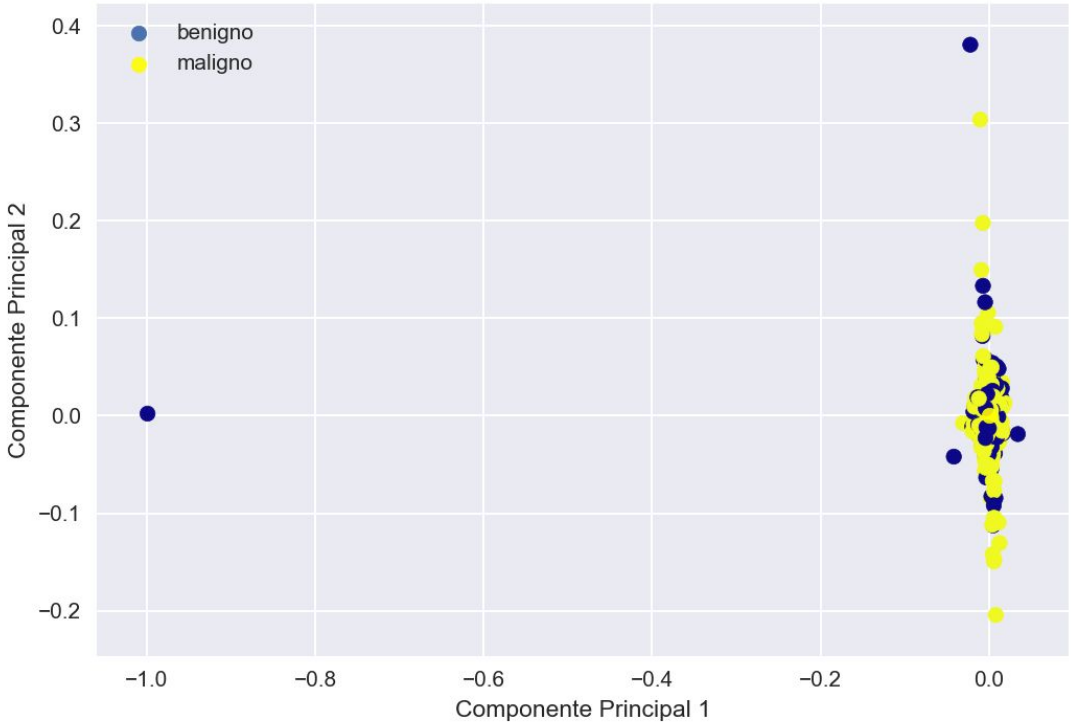

Figura 6. Componentes principais do kernel gaussiano.

Percebe-se que não é possível distinguir os alvos com o kernel gaussiano. Ao aplicar o kernel gaussiano, uma função não linear pode gerar *overfitting* em alguns dados que, no comportamento geral, pode gerar resultados discrepantes como vistos na figura 6.

Conclusão

A distinção entre os tipos de câncer de mama tem alta correlação com medidas lineares das dimensões relacionadas ao tumor, o que gera uma boa resposta para o PCA linear e uma resposta insatisfatório ao *fit* de uma função não linear. A redução por PCA apresentou resultados bem satisfatórios uma vez que reduziu de 30 dimensões para apenas 2 dimensões com separações bem evidentes entre os dois tipos de câncer.

Referências:

[1] https://saude.gov.br/saude-de-a-z/cancer-de-mama **<Disponível em 05/07/2020>**

[2] https://clinicadamama.com.br/cancer-de-mama-maligno-e-benigno/ **<Disponível em 05/07/2020>**

[3] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) **<Disponível em 05/07/2020>**

[4] https://pt.wikipedia.org/wiki/An%C3%A1lise_de_componentes_principais **<Disponível em 05/07/2020>**

[5] https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp **<Disponível em 05/07/2020>**