TP3 - Algoritmos para Bioinformática I

Lucas Resende Pellegrinelli Machado - 2018126673

Outubro 2020

1 Introdução

Com o avanço da disponibilidade de dados no mundo, foi também necessária o desenvolvimento de diversas técnicas para nos auxiliar a entender como esses dados se comportam de forma mais eficiente.

Nesse trabalho, complementando o que foi discutido nos trabalhos anteriores, apresentaremos novos meios de entender os dados disponíveis. tirar conclusões e fazer previsões sobre eles. Esse processo é de suma importância no mundo atual visto que esse tipo de problema se estende sobre praticamente todos os temas da ciência contemporânea.

Mais especificamente, trabalharemos aqui com dados que, de certa forma, se mantêm no âmbito da biologia, levando em conta características de flores, composição quimica de bebidas e informações sobre tumores no seio.

2 Entradas e saídas

Como entrada para os algoritmos utilizados, temos 3 datasets sendo deles Iris, Wine e Cancer Wisconsin. O dataset Iris é um dataset clássico para testar algoritmos de visualização de dados e consiste em 4 características de flores (mais especificamente as dimensões da pétala e da sépala) caracterizando três espécies diferentes.: Iris-Setosa, Iris-Versicolor e Iris-Virginica.

O dataset Wine é uma base de dados de classificação de vinhos onde temos 13 caracteristicas de cada vinho (como concentração de álcool, cor, intensidade da cor e outros). A ideia do dataset é que com essas características devemos classificar cada um dos vinhos entre 3 tipos de vinhos diferentes.

Já o dataset Cancer Wisconsin tem 10 valores diferentes de medições feitas nas celulas em questão como raio, textura, perimetro, area e outros. Como saida esperada do dataset temos uma classificação entre um tumor benigno e maligno.

Com esses datasets em mãos podemos fazer análises interessantes como as propostas pelo trabalho. Inicialmente uma simples visualização dos dados com dimensões reduzidas é essencial para entender os dados. Porém o primeiro resultado interessante vem com clusterizão utilizando o k-means sobre esses dados de dimensão reduzida.

Inicialmente essa análise foi feita com os dados do dataset iris, que resultou no gráfico a seguir.

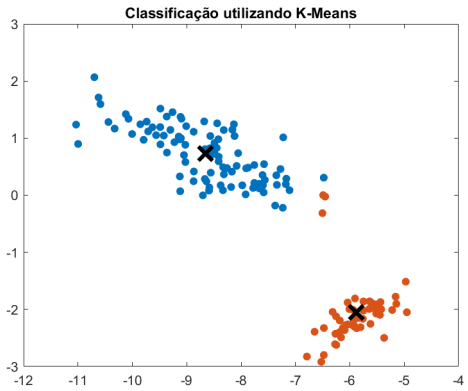

Figure 1: Clusterização em 2 grupos dos dados do dados Iris reduzidos em 2 dimensões. Nesse caso o algoritmo separou a espécie Iris-Setosa das outras.

Esse resultado é interessante, visto que podemos visualmente abstrair que de fato existem dois grupos de pontos distintos nessa figura, e o algoritmo conseguiu de maneira correta separar esses dois grupos. Essa separação clara entre os pontos mostra que esses dados são bons de serem explicados por meio desse algoritmo, o que não é algo que se aplica a todos os conjuntos de dados como veremos a seguir.

Essa mesma análise foi feita com ambos dataset Wine e Cancer Wisconsin

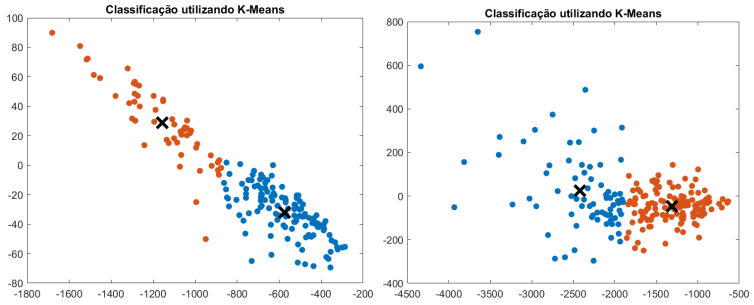

Figure 2: Clusterização em 2 grupos dos dados do Wine (na esquerda) e Cancer Wisconsin (na direita) reduzidos em 2 dimensões. Esses dados não são tão bem separáveis como o dataset anterior, pelo menos quando levamos em conta apenas os 2 autovetores com maiores autovalores.

Esses dois conjuntos de dados mostrados acima já não tem uma divisão tão clara para que o kmeans consiga fazer a separação correta entre eles no espaço R2. Em especial os dados da esquerda (que correspondem ao conjunto de dados

do câncer de seio), os dados não são nem um pouco separados de forma linear, então o algoritmo em questão tem dificuldade de conseguir separar corretamente os dados.

Outra parte do trabalho proposto consiste em criar um classificador baseado em uma regressão logística para classificar os dados de cada um dos dados aprensentados previamente.

O caso a seguir utiliza os dados da base de dados Iris, classificando de uma dada flor é da espécie Virginica ou não.

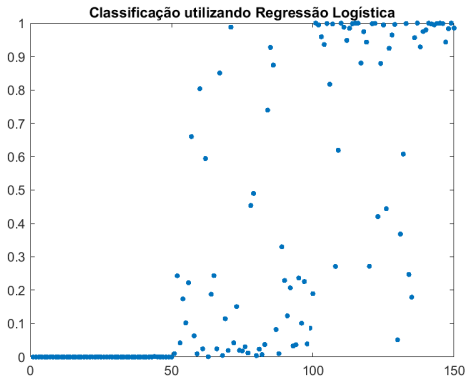

Figure 3: Regressão logística para classificar se uma flor da base de dados é do tipo Iris-Virginica ou não

É possível notar que, ao contrário do gráfico gerado em aula com essa base de dados, a regressão não é tão "limpa" pois enquanto na aula a espécie que foi isolada foi a Iris-Setosa, que é a espécie que fica, no espaço de duas dimensões, em um cluster separado, nesse trabalho a espécie isolada foi a Iris-Virginica, cujo cluster fica bem próximo ao cluser da espécie Iris-Versicolor.

O mesmo processo foi feito para os outros datasets, classificamendo os dados respectivamente em qual tipo de vinho uma instância é (no caso do dataset Wine) e se um tumor é maligno ou benigno (no caso do dataset de Câncer).

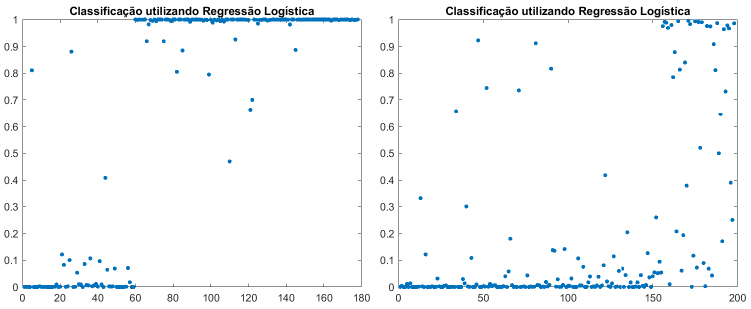

Figure 4: Regressão linear com os dados do Wine (na esquerda) para classificar se um vinho é da classe 1 ou não e Cancer Wisconsin (na direita) para classificar se um tumor é benigno ou maligno.

Pode-se notar que, denovo, o dataset do câncer de seio é relativamente mais difícil do que os outros, sendo que existem muito mais pontos menos definidos no gráfico da regressão logística.

3 Metodologia

Para representar os dados de forma matemática e que seja possível utilizar ferramentas da matemática para transforma-los, primeiramente devemos gerar o vector space model desses dados.

O vector space model utilizado foi gerado a partir das features de cada dataset, que foram transformados em uma matriz, separados da sua label. Com esses vector space models em mãos, podemos executar o algoritmo de SVD sobre essas matrizes para decompo-las em seus componentes.

Com os componentes, conseguimos descobrir quais os autovetores mais importantes para representar os dados de entrada olhando para qual o autovalor correspondente a cada um dos autovetores, sendo que aqueles com os maiores autovalores são os mais importantes. Usando os 2 autovetores com os maiores autovalores, podemos representar os dados no espaço 2D e assim clusterizar os dados utilizando o algoritmo de K-Means, que a partir de pontos iniciais aleatórios, considera cada ponto testando em qual dos clusters é melhor adiciona-lo.

Para a classificação utilizando regressão logistica, primeiro foi criado um vetor inicialmente contendo valores representando as 2 classes nas posições correspondentes às features dessas classes. Com esse vetor inicializado, podemos, a partir da relação entre os dados, podemos calcular e atualizar esse vetor de forma que que cada posição agora será a probablidade de um dado ponto estar em um dos clusters.

4 Conclusões

O trabalho constroe em cima do trabalho anterior e aprofunda os conhecimentos sobre as técnicas de clusterização e redução de dimensionalidade ao introduzir o algoritmo de classificação baseado em curvas logísticas.

Com essa ferramenta, agora podemos ter mais um nível de confiança ao criar modelos de classificação e explicação de dados, onde com dados novos, podemos com algum certeza indicar a qual classe esse dado deveria pertencer.

A aplicabilidade desses métodos também é impressionante, e como dito antes, pode ser utilizada em qualquer ramo da ciência com grande efetividade.

5 Referências bibliográficas

1. Série de vídeos "Singular Value Decomposition" por Steve Brunton.

   https://www.youtube.com/playlist?list=PLMrJAkhIeNNSVjnsviglFoY2nXildDCcv

2. Relatório de aula sobre SVD do departamento de ciência da computação da universidade de Carnegie Mellon.

   https://www.cs.cmu.edu/ venkatg/teaching/CStheory-infoage/book-chapter-4.pdf

3. Documentação do Matlab sobre o K-Means

   https://www.mathworks.com/help/stats/kmeans.html