**Trabalho Prático 2 - Visualização e Definição de Agrupamentos (Clusters)**

**Júlia Fonseca de Sena - 2018054508**

1. **Introdução**

A prática de agrupamento de dados, ou *clustering*, visa formar grupos de dados de acordo com sua semelhança, calculada a partir de uma métrica de distância a ser escolhida (no caso desse trabalho, são comparadas a euclidiana e a cosseno). Suas aplicações são diversas, como no mercado na divisão da clientela em segmentos para melhor entendimento de como se relacionam e decisões sobre posicionamento de/novos produtos, ou em análise criminológica agrupar crimes semelhantes para identificar áreas com maior densidade de um tipo de ocorrência.

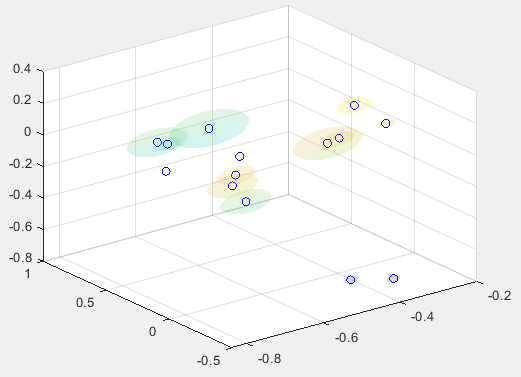
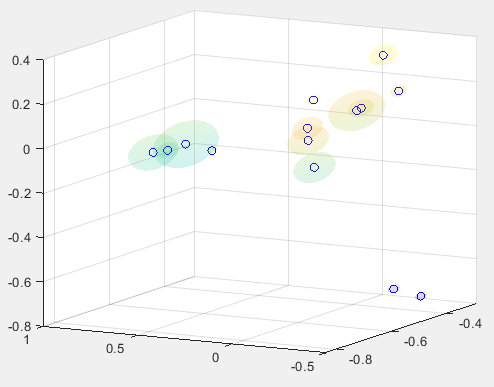
O objetivo desse trabalho é aplicar um algoritmo capaz de fazer tal agrupamento a partir de matrizes de dados originadas de experimentos, no contexto da Bioinformática. Compara-se os resultados obtidos por duas métricas diferentes, de forma a decidir qual é uma melhor escolha para o algoritmo em questão. Obtém-se, assim, uma visão prática somada a teoria já estudada.

Esse relatório possui, incluindo essa introdução, 6 seções. A seguir, apresenta-se “Entradas e saídas” que trata das necessidades do programa e do resultado final produzido. A metodologia aborda como a resolução do problema de *clustering* foi feita e a teoria do que foi implementado. Seguem as conclusões obtidas do trabalho e, por fim, as referências bibliográficas e anexos.

1. **Entradas e saídas**

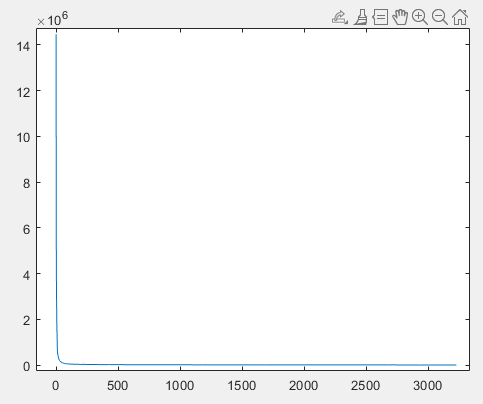
A entrada do programa é uma matriz normalizada de dados, e a saída é um gráfico com os centros de cada *cluster* computado, envolvido por uma esfera cujo diâmetro é proporcional ao número de elementos do grupo a que pertence. Por exemplo, a matriz que pode ser baixada no site <https://www.ebi.ac.uk/gxa/sc/experiments/E-MTAB-5061/downloads> pela opção “Normalised counts files (MatrixMarket archive)”. O experimento em questão trata de células pancreáticas de pessoas saudáveis e com diabetes tipo-2.

O gráfico a seguir foi resultado da execução do algoritmo com distância cosseno para a matriz no link acima. Colocou-se duas perspectivas para facilitar a análise, já que o gráfico possui 3 dimensões. Eles podem ser vistos em tamanho maior nos anexos.

1. **Metodologia**

O primeiro passo é determinar o número de agrupamentos. Faz-se isso multiplicando a matriz A (a matriz de dados citada na seção anterior) pela sua transversa (A’\*A), sendo o resultado uma matriz de combinações entre as células de A (célula x célula), e obtendo seus maiores autovalores. Plotando como no gráfico a seguir a raiz quadrada dos autovalores decrescentemente, o início da cauda longa foi utilizado para determinar o número de *clusters* (foi necessário zoom para fazer tal dedução).



A seguir, gerou-se combinações das linhas a partir da multiplicação de S e D’ (resultado da decomposição em valores singulares) a fim de obter aproximação das colunas e perceber similaridade entre elas (aproximações similares são colunas similares). A partir dessa combinação, pode-se aplica-se o algoritmo de *k-means* com os K grupos encontrados no passo anterior.

Parte-se então para o cálculo da distância e geração dos resultados. Aplicou-se, para comparação de qual métrica seria melhor, as distâncias euclidiana e cosseno. A principal diferença entre elas é que enquanto a primeira se assemelha a medir a distância entre dois pontos com uma régua, a cosseno considera a distância angular, o que tende a torna-la melhor para casos com mudança de dimensionalidade. A seguir apresenta-se as tabelas de número de elementos por grupo para euclidiana e cosseno, respectivamente:

|  |  |
| --- | --- |
| **Cluster** | **Quantidade de elementos** |
| 1 | 657 |
| 2 | 962 |
| 3 | 205 |
| 4 | 68 |
| 5 | 105 |
| 6 | 117 |
| 7 | 54 |
| 8 | 118 |
| 9 | 114 |
| 10 | 205 |
| 11 | 192 |
| 12 | 86 |
| 13 | 32 |
| 14 | 159 |

|  |  |
| --- | --- |
| **Cluster** | **Quantidade de elementos** |
| 1 | 125 |
| 2 | 428 |
| 3 | 239 |
| 4 | 321 |
| 5 | 497 |
| 6 | 172 |
| 7 | 86 |
| 8 | 93 |
| 9 | 217 |
| 10 | 195 |
| 11 | 101 |
| 12 | 61 |
| 13 | 380 |
| 14 | 314 |

1. **Conclusão**

A comparação entre os resultados das duas distâncias mostrou grande disparidade, tanto em relação há concentração de elementos em certos *clusters* (maior concentração de 962 para euclidiana e 497 para a cosseno) quanto àqueles que tiverem mais elementos (*cluster* 2 na primeira e 5 na segunda). Qual está mais correto é uma difícil conclusão sem resultados concretos para comparação, mas algumas pesquisas informaram que talvez a tridimensionalidade no programa torne distância de cosseno mais apropriada.

Por fim, foi de grande proveito observar esse algoritmo agindo sobre um exemplo real e concreto, além da teoria. Creio que foi de grande ajuda para a abstração de alguns conceitos vistos.

1. **Referências Bibliográficas**

* MACÊDO, Lincoln de. [Distância Euclidiama vs Similaridade de Cossenos](https://demacdolincoln.github.io/anotacoes-nlp/posts/distancia-euclidiama-vs-similaridade-de-cossenos/). Anotações sobre NLP, 2018. Disponível em: <https://demacdolincoln.github.io/anotacoes-nlp/posts/distancia-euclidiama-vs-similaridade-de-cossenos/> . Acesso em: 20/07/2021.
* EMMERY, Chris. Euclidian vs. Cosine Distance. 2017. Disponível em: [https://cmry.github.io/notes/euclidean-v-cosine . Acesso em 20/07/2021](https://cmry.github.io/notes/euclidean-v-cosine%20.%20Acesso%20em%2020/07/2021).

1. **Anexos**

Abaixo os gráficos resultados do algoritmo com distância cosseno, apresentados na seção 2, ampliados para maior visibilidade:

