Exercício 10: Análise de dados (parte III)

Cláudia Reis (53082), Ernesto González (52857), Ana Helena Prata (53078)

Separação de *gamers* finlandeses em grupos locais pelo método *K-Means* para 2 e 3 clusters. Separação de cores de imagens pelo método de *K-Means*. Redução de dimensionalidade em características de semicondutores com o método *PCA* e separação dos semicondutores em 3 grupos com o *K-Means*.

I. Análise de utilizadores de um jogo online na Finlândia

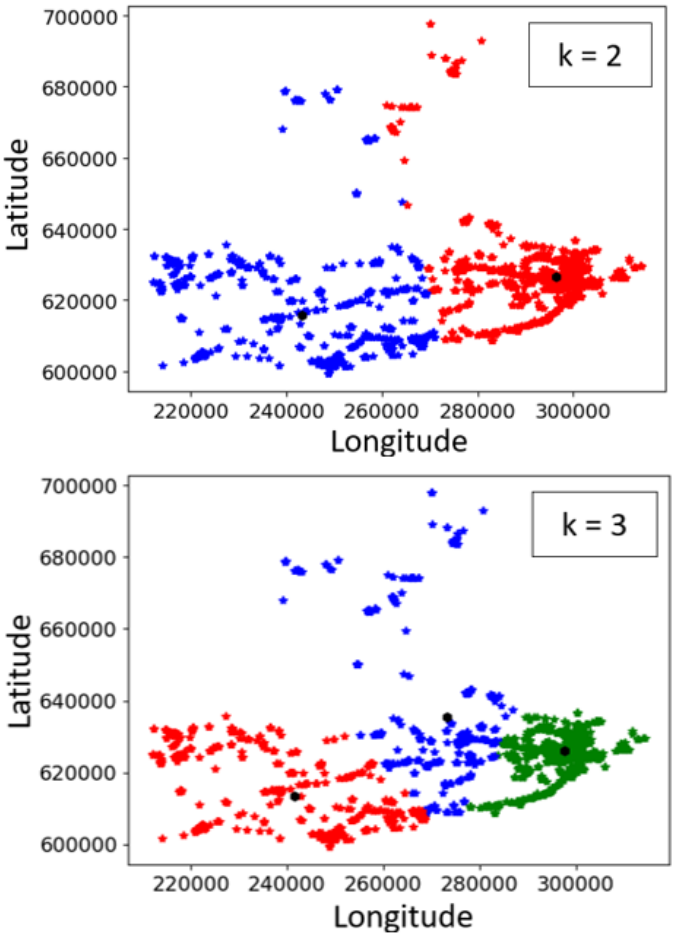

Figura 1. Representação gráfica do agrupamento dos dados fornecidos dos utilizadores de um jogo online na Finlândia em 2 e em 3 grupos geográficos para 10 iterações e centróides iniciais aleatórios. A preto encontram-se os centróides dos respetivos grupos e as diferentes cores correspondem aos diferentes grupos.

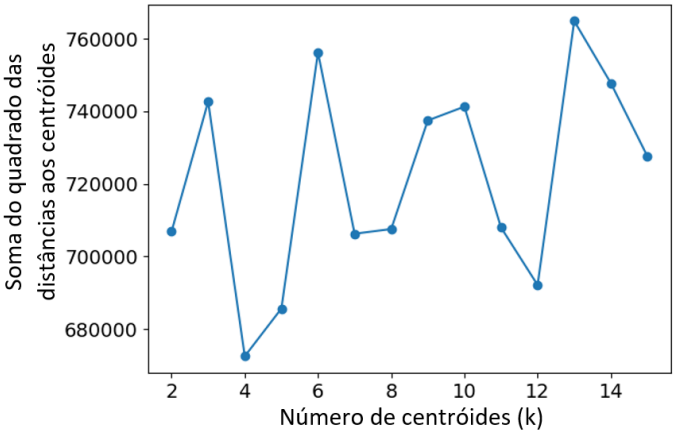

Figura 2. Representação gráfica da soma do quadrado das distâncias aos centróides em função do número de centróides (k).

Neste primeiro exercício, foram-nos fornecidos dados relativos à localização de utilizadores de um jogo online na Finlândia. Para se analisar estes dados, utilizou-se o método iterativo denominado clustering k-means, que permite agrupar por zonas de interesse estes mesmos dados consoante a sua posição num determinado espaço e, deste modo, também se consegue ver possíveis localizações para novos servidores. Este método consegue funcionar a várias dimensões, mas no âmbito deste relatório utilizou-se apenas em duas dimensões.

Neste método é necessário definir uma posição inicial para o número de centróides (k) que se quer, sendo que pode ser uma posição escolhida ou, como foi no caso deste exercício, aleatória usando o módulo *random* do *Python*. Depois de definidos os centróides, percorrem-se todos os pontos e verica-se a que centróide cada ponto pertence, sendo que pertecem ao centróide a que estiverem mais perto, formando assim grupos. Após isto, procede-se à realocação dos centróides para o "centro de massa"do grupo, isto é, o ponto médio das posições dos elementos do grupo. De seguida, verifica-se a distância de cada ponto dos dados aos centróides respetivos, formando-se novos grupos consoante o centróide mais próximo. O método vai continuar a iterar até que a posição dos centróides estabilize ou até que se atinja o número máximo de iterações definido.

Deste modo, agrupou-se os utilizadores em 2 e em 3 grupos geográficos para um máximo de 10 iterações, como se pode verificar na Figura 1. Como se pode observar, os centróides localizam-se no centro dos respetivos grupos e os grupos parecem bem definidos, pelo que se conclui que correspodem ao esperado e que o número de iterações foi adequado.

Posteriormente, traçou-se o gráfico da soma do quadrado das distâncias aos centróides em função do número de centróides (k) para assim ser possível discutir-se o número de centróides mais adequado. Observando a Figura 2, percebe-se que o número de centróides mais adequado é 4, uma vez que é o que apresenta uma menor distância dos dados aos respetivos centróides. Além disso, verifica-se ainda que a maior distância aos respetivos centróides é para 6 e 13 centróides.

II. Análise de imagens

Consideremos a Figura 3 que apresenta um composto coloidal.

Na Figura 3 podemos ver predominantemente 3 cores: um tom de vermelho, outro de azul e o fundo preto. Aplicamos o método de *K-Means* para 3 clusters para tentar separar as diferentes partes às coordenadas RGB dos pixels da imagem. A posição inicial dos centróides foi escolhida tendo em conta as componentes que queremos separar: $RGB_{avermelhado} = (111, 4, 50)$, $RGB_{azulado} = (19, 0, 90)$ e $RGB_{preto} = (9, 1, 0)$.

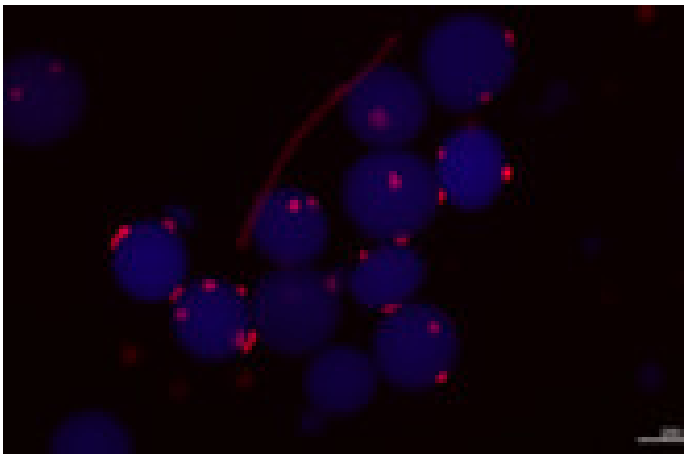

Figura 3. Composto coloidal.

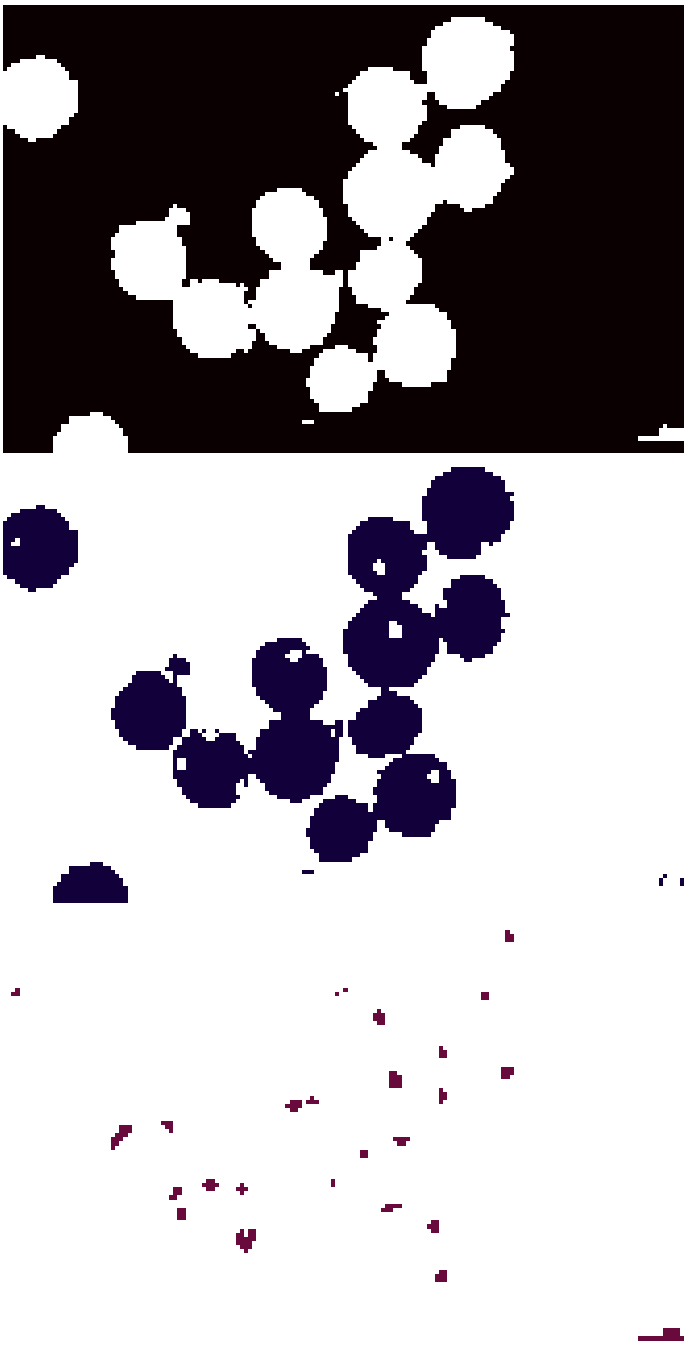

Figura 4. Separação das cores da Figura 3 pelo método *K-Means*, usando três centróides, para posições iniciais de centróides $RGB_{avermelhado} = (111, 4, 50)$, $RGB_{azulado} = (19, 0, 90)$ e $RGB_{preto} = (9, 1, 0)$ e um máximo de 15 iterações.

Agora consideremos a imagem da Figura 3 em escala de cinzentos, apresentada na Figura 5.

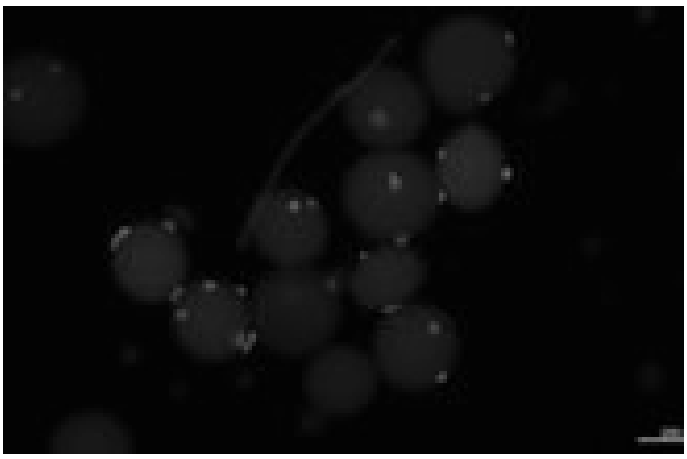

Figura 5. Imagem do composto coloidal em escala de cinzentos.

Aplicamos também à imagem do composto coloidal em escala de cinzentos (Figura 5) o método *K-Means*, com centróides 3 iniciais aleatórios e um máximo de 15 iterações. O resultado encontra-se na Figura X.

Figura 6. Separação de cores da imagem do composto coloidal em escala de cinzentos (Figura 5) pelo método de *K-Means* para 3 centróides iniciais aleatórios e um máximo de 15 iterações.

III. Redução de dimensionalidade em características de semicondutores

A partir dos dados sobre as características de vários semicondutores fornecidos num ficheiro csv foi encontrada uma 'diretriz de tendência para a variância máxima' para os dados, i.e. um vetor para o qual o valor da covariância entre cada conjunto de dados é mínimo. Isto para obter uma visualização do comportamento dos dados isoladamente (sem contar possíveis dependências que as variáveis possam ter entre si).

Para concretizar este objetivo iremos implementar o algoritmo do PCA, que apenas funciona com os dados centrados em 0 (com uma média em zero, a representação gráfica dos dados será apenas uma reflexão do comportamento da variância).

Assim, começamos por calcular a média e variância para cada uma das variáveis de estudo ($X_{atomicno}, X_{meltingpoint}, X_{VE}, X_{radii}, X_{EN}, X_{latticeconst.}$).

Iremos substituir cada ponto dos dados,X_i pelo seu correspondente normalizado em zero. Ou seja, $X_i \rightarrow \frac{X_i - \overline{X_I}}{\delta_{Xi}}$.

Calculamos também a covariância entre os dados de cada variável obtendo então uma matriz do covariância genericamente do tipo:

$$\begin{bmatrix} Var(x_1) & cov(x_1,x_2) & cov(x_1,x_3) \\ cov(x_2,x_1) & Var(x_2) & cov(x_2,x_3) \\ cov(x_3,x_1) & cov(x_3,x_2) & Var(x_3) \end{bmatrix} \quad (1)$$

Em que $\text{cov}(x_1,x_1)=\text{Var}(x_1)$,$\text{cov}(x_2,x_2)=\text{Var}(x_2)$ e $\text{cov}(x_3,x_3)=\text{Var}(x_3)$.

Como foi dito, o nosso objetivo é minimizar (se possível a zero) as relações entre os valores de variáveis diferentes ($\text{cov}(x_i,x_j\ ,\ i \neq j)$) e maximizar a variância.Ou seja, queremos obter uma matriz diagonal da forma:

$$\begin{bmatrix} Var(x_1) & 0 & 0 \\ 0 & Var(x_2) & 0 \\ 0 & 0 & Var(x_3) \end{bmatrix} \quad (2)$$

Sendo o vetor pretendido, para o qual a variância é máxima, constituído pelas entradas da diagonal principal.

Tabela I. Componentes dos vetores próprios das componente principal 1, $PC1$, e componente principal 2, $PC2$ determinados pelo PCA para os dados em semiconductors.csv. Neste problema, o espaço vetorial é formado pelas componentes (*Atomic no*, *Melting point*, *VE*, *radii*, *EN*, *lattice const.(ang)*)

Variável em estudo	**PC1**	**PC2**
Atomic no	-0.364945	-0.392646
Melting point	-0.411481	-0.089104
VE	-0.161415	-0.660326
radii	0.521120	-0.336029
EN	-0.433471	0.490540
lattice const. (ang)	-0.460413	-0.219805

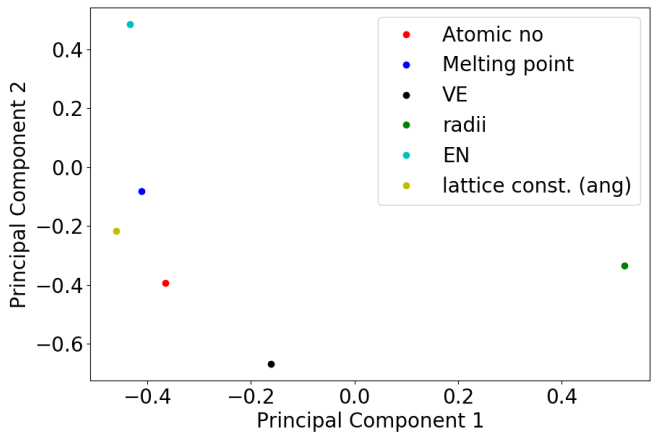

Figura 7. Componente principal 2, $PC2$, em função da componente principal 1, $PC1$, calculados pelo método PCA aplicado aos dados em semiconductors.csv.

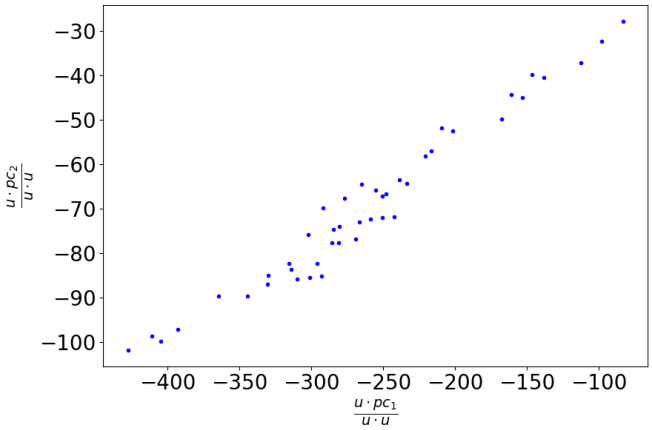

Figura 8. Escala das projeções dos dados em semiconductors.csv nos vetores diretores de $PC1$ e $PC2$. u são as coordenadas de cada semicondutor, pc_1 é o vetor diretor de $PC1$ e pc_2 o vetor diretor de $PC2$.

Aplicamos, agora o método de *K-Means* aos dados em semiconductors.csv, para 3 centróides iniciais aleatórios. Encontram-se na lista a seguir os semicondutores agrupados pelos 3 grupos encontrados:

Grupo 1: AIN, AIP, GaAs, GaSb, InSb, ZnS, ZnSe, ZnTe, CdSe, CdTe, MgS, MgSe, ZnMgS, SSMg, SSZn, ZnMgSe, ZnCdSe, SeTeZn, SeTeCd, ZnCdTe, AlGaP, PAsGa, AllnP, AsSdGa, GalnAs, PAsIn, GalnSb, AsSbln;

Grupo 2: AIAs, AISb, InAs, AlGaAs, AsSdAl, AnZnAs, AlGaSb, AllnSb;

Grupo 3: GaN, GaP, InN, InP, AlGaN, AllnN, GalnN, GalnP.