Universidade Federal do Rio de Janeiro
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

Programa de Engenharia de Sistemas e Computação

CPS863 - Aprendizado de Máquina

Prof. Dr. Edmundo de Souza e Silva
(PESC/COPPE/UFRJ)

Lista de Exercícios 2

Luiz Henrique Souza Caldas
email: lhscaldas@cos.ufrj.br

23 de outubro de 2024

Questão 1

Responda as seguintes perguntas usando o dataset fornecido:

1. Visualize os dados (em 2 ou 3 dimensões) para entender a estrutura dos dados. Explique o que você fez para visualizar as figuras. Descreva sua abordagem para visualizar os dados e quais insights podem ser obtidos do gráfico, se algum.

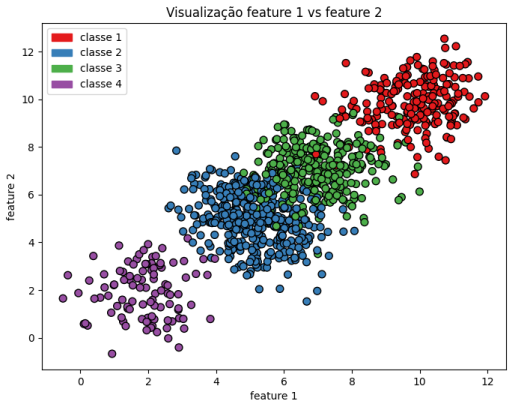
Resposta:

Para visualizar os dados, utilizei a função 'visualizar_dados', que gera gráficos em 2 e 3 dimensões a partir do dataset fornecido. No código, carreguei o dataset usando a biblioteca pandas e extraí as colunas 'feature 1', 'feature 2' e 'feature 3', além da coluna 'class label'.

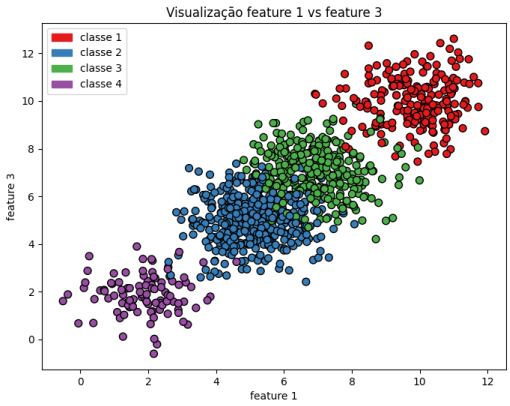
Em seguida, criei gráficos 2D para todas as combinações de features (figuras 1a, 1b e 1c). As combinações consideradas foram 'feature 1' vs 'feature 2', 'feature 1' vs 'feature 3' e 'feature 2' vs 'feature 3'. Para cada gráfico 2D, utilizei a função 'scatter' do matplotlib, onde as cores dos pontos representam as diferentes classes. Para facilitar a identificação, adicionei uma legenda correspondente a cada classe.

Por fim, também plotei os dados em um gráfico 3D (figura 1d), representando as três features simultaneamente. Esse gráfico permite uma visualização mais abrangente da estrutura dos dados.

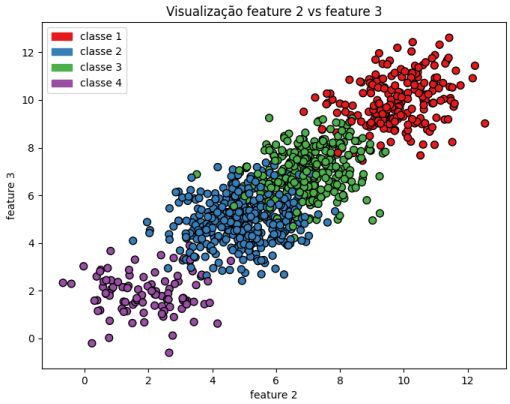
A análise visual revela que as classes parecem estar bem separadas, principalmente no gráfico 3D, o que sugere que um modelo de mistura de gaussianas pode ser adequado para representar os dados. As classe 1 e 4, em todas as vistas, estão bem afastadas das demais classes, enquanto as classes 2 e 3 possuem uma fronteira mais tenue entre si. Existem alguns pontos de sobreposição entre as classes, mas no geral elas estão bem separadas.



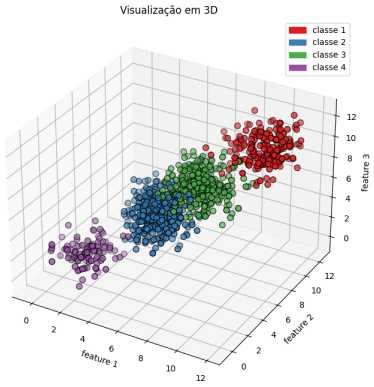
(a) Feature 1 vs Feature 2



(b) Feature 1 vs Feature 3



(c) Feature 2 vs Feature 3



(d) Gráfico 3D das features

Figura 1: Visualização dos dados em 2D e 3D

2. Ajuste uma mistura de gaussianas com 4 componentes ao conjunto de dados. Calcule TODOS os parâmetros necessários para o modelo. Explique todas as etapas. Forneça detalhes de como você determina os parâmetros de melhor ajuste para cada modelo de mistura e descreva o processo de ajuste do modelo.

Resposta:

Passos para o MLE em uma Mistura de Gaussianas com Labels Conhecidos:

(a) **Função de Verossimilhança:**
Dado que temos os rótulos y_i que indicam a qual gaussiana cada ponto x_i pertence, a função de verossimilhança é dada por:

$$L(\mu_k, \Sigma_k, \pi_k) = \prod_{i=1}^{N} \prod_{k=1}^{K} [\pi_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)]^{\mathbb{1}(y_i=k)}$$

Onde $\mathbb{1}(y_i = k)$ é a função indicadora que vale 1 quando o dado x_i pertence à classe k e $\mathcal{N}(x_i|\mu_k, \Sigma_k)$ é a densidade da gaussiana multivariada.

(b) **Função de Log-Verossimilhança:**
Tomamos o logaritmo da função de verossimilhança para simplificar o processo de maximização:

$$\log L(\mu_k, \Sigma_k, \pi_k) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}(y_i = k) \left[\log \pi_k - \tfrac{1}{2} \log |2\pi\Sigma_k| - \tfrac{1}{2}(x_i - \mu_k)^\top \Sigma_k^{-1}(x_i - \mu_k)\right]$$

(c) **Maximização via Derivação:**

- **Derivada em relação a μ_k:**
A derivada da log-verossimilhança em relação a μ_k é:

$$\frac{\partial \log L}{\partial \mu_k} = \sum_{i=1}^{N} \mathbb{1}(y_i = k)\Sigma_k^{-1}(x_i - \mu_k)$$

Igualando a zero, obtemos a estimativa para μ_k:

$$\mu_k = \frac{1}{N_k} \sum_{x_i \in X_k} x_i$$

Onde N_k é o número de amostras na classe k.

- **Derivada em relação a Σ_k:**
A derivada da log-verossimilhança em relação à matriz de covariância Σ_k é:

$$\frac{\partial \log L}{\partial \Sigma_k} = \sum_{i=1}^{N} \mathbb{1}(y_i = k) \left(\frac{1}{2}\Sigma_k^{-1} - \frac{1}{2}\Sigma_k^{-1}(x_i - \mu_k)(x_i - \mu_k)^\top \Sigma_k^{-1}\right)$$

Igualando a zero, obtemos a estimativa para Σ_k:

$$\Sigma_k = \frac{1}{N_k} \sum_{x_i \in X_k} (x_i - \mu_k)(x_i - \mu_k)^\top$$

Resposta (continuação):

(c) **Maximização via Derivação (cont.):**

- **Derivada em relação a π_k:**
  Com a restrição $\sum_{k=1}^{K} \pi_k = 1$, derivamos a log-verossimilhança em relação a π_k:

$$\frac{\partial \log L}{\partial \pi_k} = \sum_{i=1}^{N} \frac{\mathbb{1}(y_i = k)}{\pi_k}$$

  Igualando a zero, obtemos a estimativa para π_k:

$$\pi_k = \frac{N_k}{N}$$

(d) **Resumo das Soluções Fechadas:**
- Médias: $\mu_k = \frac{1}{N_k} \sum_{x_i \in X_k} x_i$
- Matrizes de Covariância: $\Sigma_k = \frac{1}{N_k} \sum_{x_i \in X_k} (x_i - \mu_k)(x_i - \mu_k)^\top$
- Pesos: $\pi_k = \frac{N_k}{N}$

Implementação: Utilizando um código Python (presente no repositório no final deste relatório), as fórmulas finais para cada parâmetro foram implementadas e os resultados podem ser vistos nas tabelas 1, 2, 3 e 4.

Tabela 1: Médias das Gaussianas

Classe	Médias
1	[9.8397, 9.8468, 9.9668]
2	[5.1378, 4.9909, 5.0114]
3	[6.9380, 7.0204, 7.0432]
4	[1.9207, 1.9505, 1.9004]

Tabela 2: Covariâncias das Gaussianas

Classe	Matrizes de Covariância
1	$\begin{bmatrix} 1.0345 & 0.2227 & 0.0477 \\ 0.2227 & 1.0632 & 0.2896 \\ 0.0477 & 0.2896 & 1.0639 \end{bmatrix}$
2	$\begin{bmatrix} 0.9515 & -0.2641 & 0.0375 \\ -0.2641 & 1.1534 & 0.2035 \\ 0.0375 & 0.2035 & 0.9640 \end{bmatrix}$
3	$\begin{bmatrix} 0.9995 & 0.1084 & -0.1126 \\ 0.1084 & 0.9665 & 0.2504 \\ -0.1126 & 0.2504 & 0.9958 \end{bmatrix}$
4	$\begin{bmatrix} 0.9279 & 0.1247 & 0.0622 \\ 0.1247 & 1.2607 & 0.0047 \\ 0.0622 & 0.0047 & 0.6999 \end{bmatrix}$

Tabela 3: Pesos das Gaussianas

Classe	Peso
1	0.2
2	0.4
3	0.3
4	0.1

Tabela 4: Log-Verossimilhança das Gaussianas

Classe	Log-LH
1	-849.8800
2	-1701.1983
3	-1275.4746
4	-424.0718

3. Suponha que as classes 1 e 2 sejam uma mesma classe. Ajuste uma mistura de gaussianas com 3 componentes ao conjunto de dados.

Resposta:

Para responder a essa pergunta, foram feitas as seguintes alterações no código do item anterior:

- Unificação das Classes: A classe 2 foi substituída por 1 na coluna 'class label' para tratar classes 1 e 2 como uma única classe.
- Ajuste da Mistura de Gaussianas: A mistura de gaussianas foi ajustada utilizando 3 componentes, refletindo a unificação das classes.

Os resultados obtidos estão nas tabelas 5, 6, 7 e 8.

Tabela 5: Médias das Gaussianas

Classe	Médias
1	[6.7051, 6.6095, 6.6632]
2	[5.1378, 4.9909, 5.0114]
3	[6.9380, 7.0204, 7.0432]
4	[1.9207, 1.9505, 1.9004]

Tabela 6: Covariâncias das Gaussianas

Classe	Matrizes de Covariância
1	$\begin{bmatrix} 5.8987 & 4.9804 & 5.2273 \\ 4.9804 & 6.3703 & 5.5882 \\ 5.2273 & 5.5882 & 6.4617 \end{bmatrix}$
2	$\begin{bmatrix} 0.9515 & -0.2641 & 0.0375 \\ -0.2641 & 1.1534 & 0.2035 \\ 0.0375 & 0.2035 & 0.9640 \end{bmatrix}$
3	$\begin{bmatrix} 0.9995 & 0.1084 & -0.1126 \\ 0.1084 & 0.9665 & 0.2504 \\ -0.1126 & 0.2504 & 0.9958 \end{bmatrix}$
4	$\begin{bmatrix} 0.9279 & 0.1247 & 0.0622 \\ 0.1247 & 1.2607 & 0.0047 \\ 0.0622 & 0.0047 & 0.6999 \end{bmatrix}$

Tabela 7: Pesos das Gaussianas

Classe	Peso
1	0.6
2	0.4
3	0.3
4	0.1

Tabela 8: Log-Verossimilhança das Gaussianas

Classe	Log-LH
1	-2553.9509
2	-1701.1983
3	-1275.4746
4	-424.0718

4. Suponha que as classes 1, 2 e 3 sejam uma mesma classe. Ajuste uma mistura de gaussianas com 2 componentes ao conjunto de dados.

Resposta:

Para responder a essa pergunta, fiz as seguintes alterações no código do item 2:

- Unificação das Classes: As classes 3 e 2 foram substituída por 1 na coluna 'class label' para tratar classes 1, 2 e 3 como uma única classe.
- Ajuste da Mistura de Gaussianas: A mistura de gaussianas foi ajustada utilizando 2 componentes, refletindo a unificação das classes.

Os resultados obtidos estão nas tabelas 9, 10, 11 e 12.

Tabela 9: Médias das Gaussianas

Classe	Médias
1	[6.7827, 6.7465, 6.7899]
2	[5.1378, 4.9909, 5.0114]
3	[6.9380, 7.0204, 7.0432]
4	[1.9207, 1.9505, 1.9004]

Tabela 10: Matrizes de Covariância das Gaussianas

Classe	Matrizes de Covariância
1	$\begin{bmatrix} 4.2748 & 3.3758 & 3.4652 \\ 3.3758 & 4.6035 & 3.8414 \\ 3.4652 & 3.8414 & 4.6687 \end{bmatrix}$
2	$\begin{bmatrix} 0.9515 & -0.2641 & 0.0375 \\ -0.2641 & 1.1534 & 0.2035 \\ 0.0375 & 0.2035 & 0.9640 \end{bmatrix}$
3	$\begin{bmatrix} 0.9995 & 0.1084 & -0.1126 \\ 0.1084 & 0.9665 & 0.2504 \\ -0.1126 & 0.2504 & 0.9958 \end{bmatrix}$
4	$\begin{bmatrix} 0.9279 & 0.1247 & 0.0622 \\ 0.1247 & 1.2607 & 0.0047 \\ 0.0622 & 0.0047 & 0.6999 \end{bmatrix}$

Tabela 11: Pesos das Gaussianas

Classe	Peso
1	0.9
2	0.4
3	0.3
4	0.1

Tabela 12: Log-Verossimilhança das Gaussianas

Classe	Log-LH
1	-3830.7946
2	-1701.1983
3	-1275.4746
4	-424.0718

5. Mostre como estimar qual dos 3 modelos (2, 3 ou 4 gaussianos) melhor representa o conjunto de dados original, ignorando as classes do modelo. Justifique como chegou ao resultado, usando técnicas de avaliação de modelos como AIC, BIC ou outro critério. Estude e explique o que é AIC e BIC.

Resposta:

Definições de AIC e BIC:

- **AIC (Critério de Informação de Akaike)**: O AIC é definido pela fórmula:

$$\text{AIC} = 2k - 2\ln(L)$$

  onde k é o número de parâmetros do modelo e L é a função de verossimilhança do modelo ajustado. O AIC penaliza a complexidade do modelo, favorecendo aqueles que oferecem um bom ajuste com menos parâmetros.

- **BIC (Critério de Informação Bayesiano)**: O BIC é calculado pela fórmula:

$$\text{BIC} = \ln(n)k - 2\ln(L)$$

  onde n é o número de observações, k é o número de parâmetros do modelo, e L é novamente a função de verossimilhança. O BIC penaliza mais severamente modelos complexos, especialmente quando n é grande, o que pode levar a uma preferência por modelos mais simples.

Utilização dos Critérios:
Para determinar qual dos modelos de mistura gaussiana (2, 3 ou 4 componentes) melhor representa o conjunto de dados original, calculamos os critérios AIC e BIC para cada modelo ajustado. Os resultados estão resumidos na Tabela 13.
Analisando os valores de AIC e BIC, podemos identificar o modelo com o menor valor, que indica o melhor ajuste aos dados, levando em consideração a penalização pela complexidade do modelo. Se as diferenças nos valores de AIC e BIC forem pequenas, poderíamos preferir o modelo com menos componentes devido à sua simplicidade, o que é uma prática comum na seleção de modelos.
Neste exercício, o modelo com menos componentes (2 componentes) também apresentou os menores valores de AIC e BIC, indicando que é o modelo preferido para representar o conjunto de dados original.

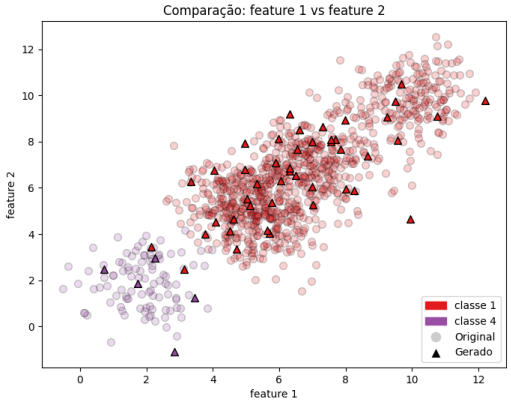
Tabela 13: AIC e BIC para Modelos de 2, 3 e 4 Componentes

Modelo	AIC	BIC
2 componentes	8523.73	8558.09
3 componentes	8528.99	8582.98
4 componentes	8531.25	8604.87

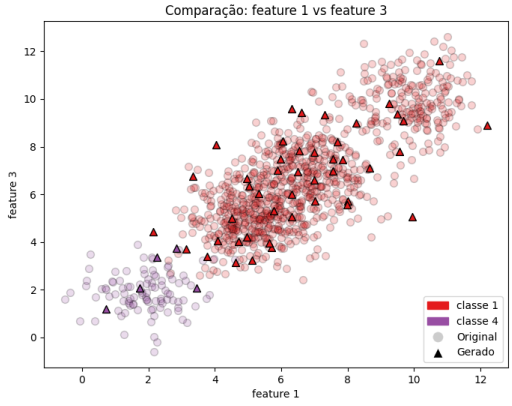
6. Gere novas amostras com base no melhor modelo pelo seu critério. Plote os resultados. Visualize as amostras geradas e compare-as com o conjunto de dados original.

Resposta:

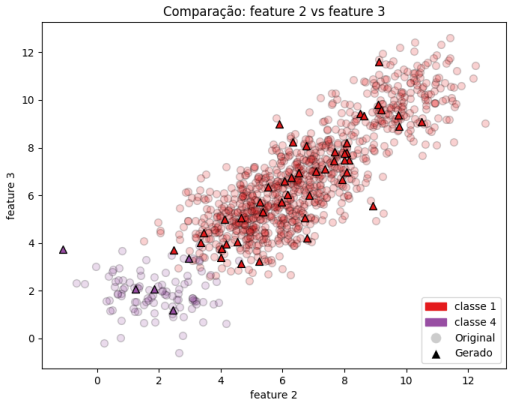
Para gerar novas amostras, utilizamos o melhor modelo de mistura de gaussianas, que identificou três classes a partir dos dados originais. As amostras geradas foram obtidas a partir dos parâmetros estimados do modelo, incluindo médias, covariâncias e pesos. Em seguida, realizamos a visualização das amostras geradas em comparação com os dados originais, utilizando diferentes marcadores para diferenciá-los: círculos para os dados originais e triângulos para as amostras geradas. A figura 2 ilustra os resultados, permitindo uma análise visual clara das diferenças e semelhanças entre os conjuntos de dados.



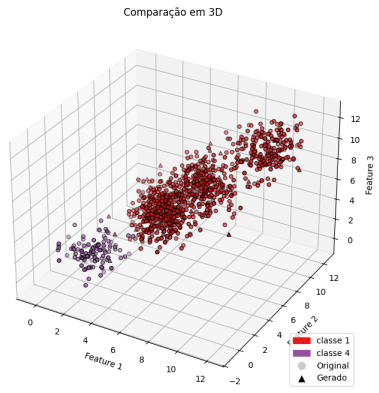
(a) Feature 1 vs Feature 2



(b) Feature 1 vs Feature 3



(c) Feature 2 vs Feature 3



(d) Gráfico 3D das features

Figura 2: Visualização dos dados em 2D e 3D

Questão 2

É dado um novo dataset com vários dados faltantes.

1. Descreva as etapas para executar a imputação de dados para as amostras incompletas. Explique sua abordagem. Descreva (mostre a matemática) que você usou para preencher os valores ausentes.

Previsão da Feature Faltante

Passo a Passo para Calcular $\hat{x}_1$

1. Definição da Expectativa Condicional: A previsão de x_1 dada x_2, x_3 e y é expressa como a expectativa condicional:

$$\hat{x}_1 = \mathbb{E}[x_1|x_2, x_3, y]$$

2.*Distribuição Condicional: Utilizamos a distribuição condicional em termos da distribuição conjunta:

$$p(x_1|x_2, x_3, y) = \frac{p(x_1, x_2, x_3|y)}{p(x_2, x_3|y)}$$

3. Distribuição Conjunta: A distribuição conjunta pode ser modelada como uma mistura de gaussianas:

$$p(x_1, x_2, x_3|y) = \sum_{k=1}^{K} \pi_k \mathcal{N}([x_1, x_2, x_3]|\mu_k^y, \Sigma_k^y)$$

4. Expectativa Condicional: A previsão condicional $\hat{x}_1$ pode ser expressa como:

$$\hat{x}_1 = \frac{1}{p(x_2, x_3|y)} \sum_{k=1}^{K} \pi_k \mathbb{E}[x_1|x_2, x_3, k]$$

5. Fórmula Final da Previsão: A previsão $\hat{x}_1$ resulta na fórmula:

$$\hat{x}_1 = \sum_{k=1}^{K} \pi_k \left(\mu_k^y + \Sigma_{12}\Sigma_{kk}^{-1}(x_{\text{known}} - \mu_{k,\text{known}})\right)$$

onde:

- Σ_{12} é a covariância entre x_1 e as features conhecidas x_2 e x_3.
- Σ_{kk} é a covariância entre as features conhecidas.

Reposta(continuação):

Analogamente, o cálculo de $\hat{x}_2$ e $\hat{x}_3$ é feito de forma similar.

$$\hat{x}_2 = \sum_{k=1}^{K} \pi_k \left(\mu_k^y + \Sigma_{21} \Sigma_{kk}^{-1} (x_{\text{known}} - \mu_{k,\text{known}}) \right)$$

$$\hat{x}_3 = \sum_{k=1}^{K} \pi_k \left(\mu_k^y + \Sigma_{31} \Sigma_{kk}^{-1} (x_{\text{known}} - \mu_{k,\text{known}}) \right)$$

2. Execute os cálculos necessários para imputar os valores ausentes. Forneça os detalhes matemáticos e demonstre o processo de imputação de valores ausentes para o conjunto de dados fornecido.

Resposta:

Para preencher os valores faltantes no conjunto de dados, foi utilizada a previsão condicional com mistura de gaussianas. As etapas realizadas são as seguintes:

(a) Carregamento do Dataset: O conjunto de dados foi carregado utilizando a biblioteca Pandas, onde as colunas 'feature 2' e 'feature 3' foram convertidas para o tipo numérico, tratando valores não numéricos como NaN. Esta conversão é crucial para garantir que todos os dados sejam adequadamente interpretados.

(b) Identificação de Linhas Completas e Incompletas: Foram separadas as linhas que contêm dados completos (sem NaN) e as que possuem pelo menos um valor ausente. As linhas completas foram usadas para ajustar o modelo de mistura de gaussianas.

(c) Ajuste do Modelo de Mistura de Gaussianas: Utilizando as linhas completas, o modelo foi ajustado para estimar os parâmetros necessários para prever os valores ausentes. A função `ajustar_mistura_gaussianas` foi utilizada para essa tarefa.

(d) Preenchimento dos Valores Faltantes: Para cada linha com dados faltantes, foi identificada a coluna com valor ausente. Os valores conhecidos da linha foram utilizados para prever o valor faltante, usando a função `calcular_previsao_condicional` com base nos resultados do modelo ajustado. O valor previsto foi então preenchido no DataFrame. As linhas que foram preenchidas podem ser observadas na tabela 14.

feature 1	**feature 2**	**feature 3**	**class label**
2.839190300459528	9.56464817530692	7.750530975334812	3
4.492978036403656	4.595353306087147	0.791578947368421	2
5.8733673095703125	7.17497368316246	0.8905263157894735	3
6.9206104800105095	2.2171797008249507	7.46922239639611	3

6.694144755601882	4.938487235652284	0.8905263157894735	3
6.708944767713547	2.109366426414244	7.106021649267843	3
1.6811296929918367	4.2475351285698	4.79874519211732	2
9.936176501214504	2.04472438940024	10.01282561819705	1
6.315349519252777	1.790139209072315	6.030610810704607	3
1.748088993363427	5.888952282607289	7.447549000035647	3
1.929783437874051	9.449970443197673	10.207371287821273	1
3.7911490201950073	2.3460556913090187	5.927534326445659	2
2.0301167494036134	9.941293360481612	9.911052952029722	1
4.348752379417419	1.713019560182466	4.328107931312082	2
1.9578048353556075	9.587188626741376	10.931736145893224	1
2.441918847473664	8.22632235851057	8.098441316210339	3
2.151514055226405	5.436006256556076	6.359691261031645	2
4.972970107570291	2.5725555649798286	6.499807943433078	2
1.9677930111819368	4.971817448465	4.442982404013302	2
6.072207152843475	1.5208369022000923	5.123386727269814	3
3.144804000854492	1.3456178903579712	0.4126315789473684	4
7.804545402526855	6.086988417393712	0.8905263157894735	3
1.9602810542679305	4.952837770091846	5.28833540581244	2
7.74591988325119	8.084379324586928	0.8905263157894735	3
1.6114837256483783	4.071567923845636	5.165193804172858	2
6.761785760521889	2.0930729759149465	7.051132365670919	3
2.6984834671020503	7.244964335395697	0.791578947368421	2
5.453582972288132	1.5478840975311094	3.91087737408126	2
2.0641982679167863	5.2153945598961355	5.62621040266761	2
3.691768765449524	2.191045305439699	5.535885745126899	2
2.2119334086921807	7.451548717225432	6.228367138081465	3
1.8495722269951174	6.230828424274331	7.427046872055546	3
5.461026310920715	2.0015197729370575	5.057031341197352	2
3.5419803857803345	2.4048383989315463	6.076054465385556	2
5.071092829108238	4.28863785920351	0.791578947368421	2
6.691092014312744	1.631090555329608	4.121106456284914	2
1.9677491884476326	6.628942301507982	6.117473010913241	3
7.624812126159668	6.561308513552811	0.8905263157894735	3
6.777165666222572	1.7062844042569676	5.748121220014607	3
2.684110058764383	6.781661052729159	4.309300230321324	2
4.048401355743408	0.18524008267804193	1.795694679021835	4
8.806843280792236	9.267883697468749	0.20421052631578948	1
4.558074802160263	2.2311184042384125	5.637134266027904	2
5.613296031951904	2.0494108785785947	5.178032804919321	2
2.3574528075703287	5.9563302318931175	6.079464032391152	2
5.156202197074889	2.6747955473839182	6.758127047911496	2
9.117786347866058	1.9714560538475097	9.65403737708832	1
6.741346657276154	7.449426608107276	0.8905263157894735	3
3.685867547988892	6.881593746400178	0.791578947368421	2

5.406050622463225	1.8005535212720136	4.549270864915992	2

Tabela 14: Conjunto de dados com valores faltantes preenchidos (apenas as linhas que foram preenchidas).

Codigos

Os códigos utilizados para a resolução dos exercícios estão disponíveis no repositório do GitHub: `https://github.com/lhscaldas/cps863/`