Relatório 2 : Análise Multivariada 2

Tharso Monteiro

7 de dezembro de 2021

1 Resumo

Neste relatório, será exposta a utilização das técnicas de agrupamento (*Clustering*, em inglês) e análise discriminante (*Discriminant Dnalysis*).

2 Análise de agrupamento

Para aplicação da análise de agrupamento, foi utilizado o conjunto de dados *Credit Card Dataset for Clustering*, disponível em `https://www.kaggle.com/arjunbhasin2013/ccdata`.

Este conjunto de dados contém informações de 9000 usuários ativos de cartões de crédito durante os últimos seis meses antes da coleta dos dados. As variáveis contidas no conjunto descrevem, a frequência de compras e empréstimos feitos com o cartão de débito e crédito, limite, saldo, frequência de atualizações do saldo, entre outras informações. A descrição completa das variáveis está disponível no link acima.

Importamos os dados. Verificamos que há 313 valores ausentes na variável MINIMAL_PAYMENTS, que descreve o menor pagamento feito pelo usuário. Como essa variável possui valores extremos, imputamos os valores ausentes com a mediana.

Também havia uma observação cuja variável CREDIT_LIMIT, que descreve o limite do cartão de crédito, estava ausente. Também imputamos este valor com a mediana.

2.1 Agrupamento K-Means

O algoritmo K-means exige que um número fixo de *clusters* seja definido *a priori*. Como não temos uma ideia prévia de quantos grupos serão utilizados para classificar os tipos de usuários de cartão, iremos utilizar um gráfico *scree*, que ilustra a soma de quadrados dentro do agrupamento (WCSS), por número de agrupamentos. O gráfico está disponível na Figura 1.

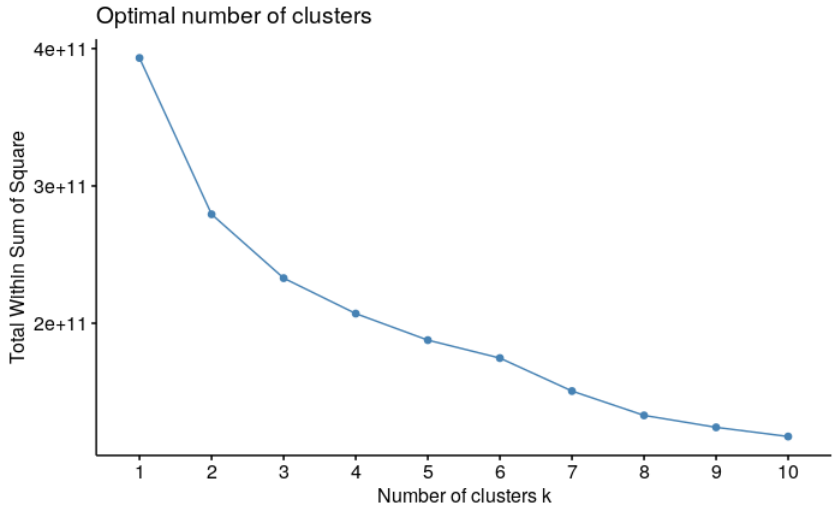

Figura 1: WCSS por número de *clusters*

Vê-se que até 3 *clusters*, há uma redução significativa do WCSS. Um aumento excessivo do número de agrupamentos não é recomendado, pois pode causar *overfitting*.

Utilizando o algoritmo K-means, realizamos o agrupamento em 3 *clusters*. Um gráfico ilustrando a categorização dos grupos pelos dois primeiros componentes principais está disponível da Figura 2.

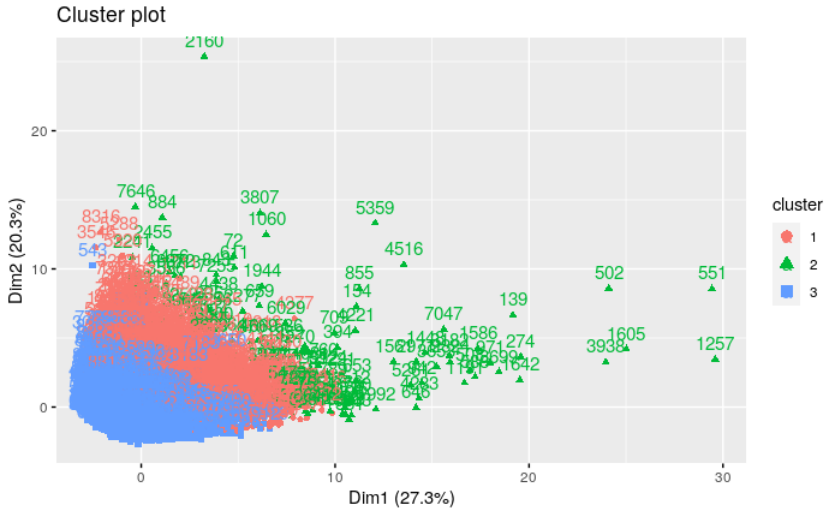

Figura 2: Agrupamento de acordo com os 2 primeiros CPs

Analisando os componentes principais, vemos que o primeiro componente atribui peso maior às variáveis referentes ao número e frequência de compras, enquanto o segundo atribui peso maior ao valor e frequência de empréstimos.

Deste modo, podemos atribuir a seguinte interpretação aos agrupamentos :

- Agrupamento 1 : Usuários não-frequentes do cartão de crédito.
- Agrupamento 2 : Usuários moderados do cartão de crédito.
- Agrupamento 3 : Usuários severos do cartão de crédito.

2.2 Agrupamento Hierárquico

Para a aplicação do agrupamento hierárquico, utilizamos o conjunto de dados *Economic Data and Statistics on World Economy and Economic Freedom*, da fundação Heritage.

Esse conjunto de dados contém informações sobre indicadores econômicos e índices ligados à liberdade econômica em 173 países.

Após importação e limpeza dos dados, realizamos uma análise de agrupamento hierárquico. O dendograma resultante está ilustrado na Figura 3.

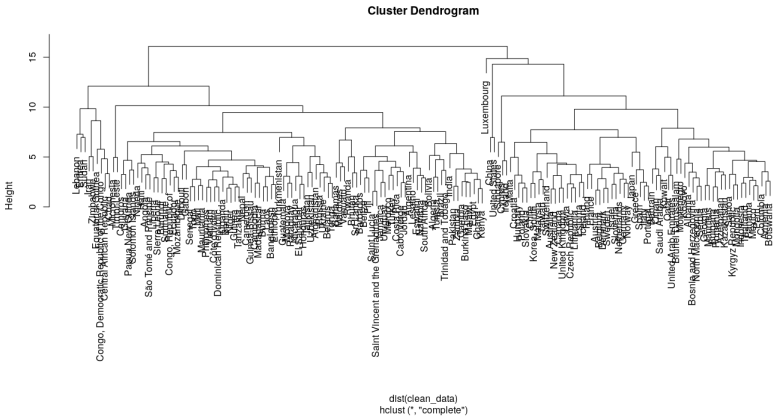

Figura 3: Dendograma da análise de agrupamento hierárquica

Em seguida, "cortamos"o dendograma de maneira que resulte em 5 agrupamentos, para facilitar a interpretação e evitar *overfitting*.

Em seguida, plotamos os dois componentes principais referentes a cada país, separando-os de acordo com seu grupo atribuído (Figura 4. Os nomes dos países foram omitidos para facilitar a visualização.

O primeiro componente principal atribui peso positivo maior às variáveis que indicam maior liberdade financeira, de negócios, investimentos e maior PIB *per capita*. O segundo componente principal atribui pesos negativos em alto valor absoluto para as variáveis referentes aos impostos. Isto é, países com valores maiores da segunda componente principal têm taxas de imposto mais baixas.

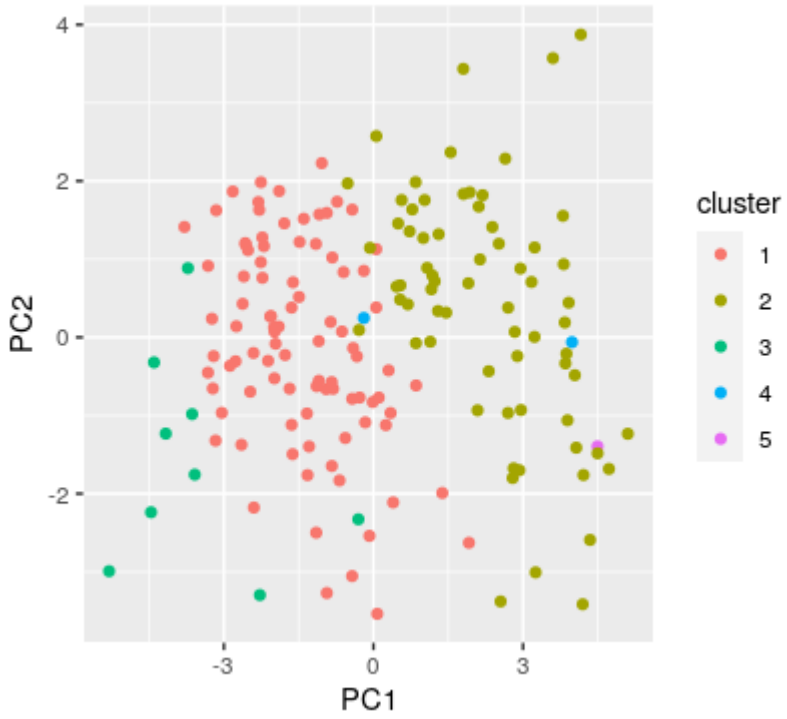

Figura 4: Agrupamento por componentes principais

Como o número de observações é grande, não iremos detalhar a classificação individual de cada país. Entretanto, os *clusters* podem ser interpretados da seguinte maneira:

- Cluster 1 : Países subdesenvolvidos e em desenvolvimento. Neste grupo se encontram a maior parte dos países do sudeste asiático, América Latina e África, assim como a Índia e Rússia. Este é o agrupamento mais numeroso.
- Cluster 2 : Países desenvolvidos que não são potências globais. Nesse grupo se encontram a maioria dos países europeus, Austrália, Nova Zelândia, e países da Ásia e América Latina com maiores índices de liberdade econômica, como Taiwan, Japão, Panamá, Chile, México e Peru.
- Cluster 3 : Países em conflito ou em grave crise econômica. Nesse grupo se encontram : República do Chade, Eritreia, Irã, Sudão, Zimbábue, Líbano, entre outros.
- Cluster 4 : Potências mundiais. Nesse agrupamento, se encontram somente a China e os Estados Unidos.
- Cluster 5 : Neste agrupamento, se encontra apenas Luxemburgo, um pequeno país europeu com altíssimos níveis de liberdade econômica.

2.3 Agrupamento Baseado em Misturas

Para a demonstração do agrupamento baseado em misturas, utilizamos o conjunto de dados **banknote**, contido no pacote **mclust**. Esse conjunto contém dados de seis medições feitas em 100 notas verdadeiras e 100 notas falsas de mil francos suíços.

Utilizando o pacote citado acima, removemos a variável referente à genuinidade das notas do conjunto de dados e realizamos o agrupamento baseado em misturas de Gaussianas para dois *clusters*. O modelo escolhido foi o VVE, em que as variáveis normais possuem volume variável, forma variável e orientação igual.

A classificação em cada agrupamento por pares de variáveis está ilustrada na Figura 5.

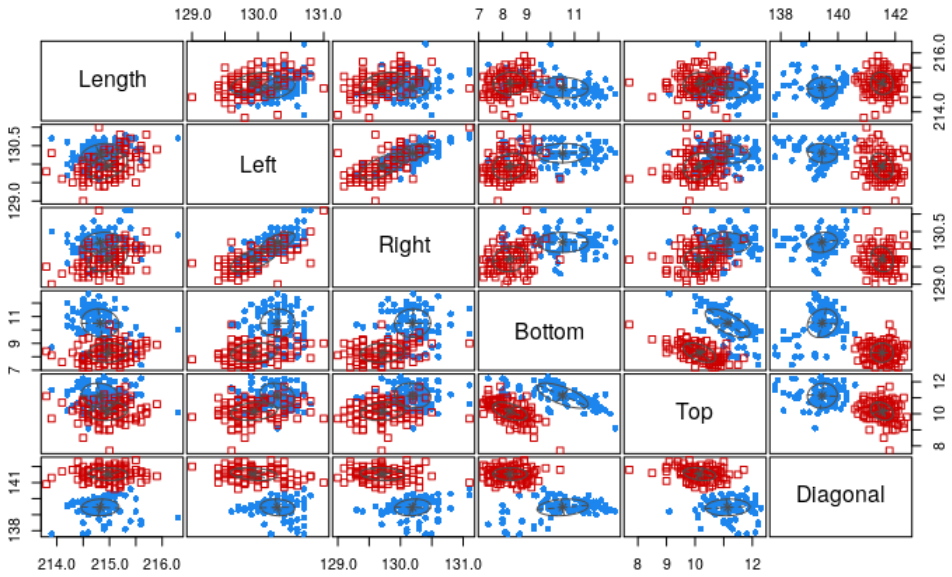

Figura 5: Agrupamento por pares de variáveis no modelo de misturas

A matriz de confusão está disponível abaixo. Vê-se que o agrupamento obteve um resultado bastante satisfatório, com apenas um erro.

	Agrupamento 1	Agrupamento 2
Falsificação	100	0
Genuíno	1	99

Tabela 1: Matriz de confusão do agrupamento por misturas

3 Análise Discriminante

Aplicamos a análise discriminante linear no mesmo conjunto de dados utilizado para a aplicação do agrupamento baseado em misturas. Desta vez, não excluímos a variável contendo informações sobre a autenticidade das notas, pois a análise discriminante é uma técnica de aprendizado supervisionado.

Separamos o conjunto de dados em um de treino e um de teste.

Os scores de cada grupo estão ilustrados na Figura 6.

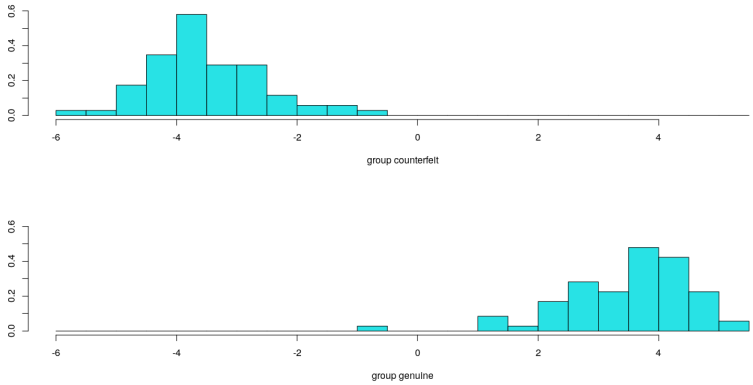

Figura 6: Histograma por grupo : Análise Discriminante Linear

A matriz de confusão para o conjunto de teste, com 60 observações, está descrita abaixo:

	Agrupamento 1	Agrupamento 2
Falsificação	31	0
Genuíno	0	29

A análise discriminante linear, assim como o agrupamento baseado em misturas, resulta num agrupamento bastante satisfatório.

Referências