UERJ - UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO
IME - INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
CComp - Programa de Pós-Graduação em Ciências Computacionais
IME12566 Tópicos Especiais: Inteligência Artificial

Prof.[a] Vinicius Layter Xavier

Relatório n[o]: k-means e ridge regression em Customer Mall dataset

Aluno: Guilherme Fontes dos Reis
e-mail: guilherme.comp.uerj@gmail.com

3 de junho de 2021

1 Objetivo

O objetivo deste trabalho é executar métodos apresentados no curso em algum dataset público. Para este trabalho foi escolhido o dataset público de Mall Customers[1] e o método de classificação não supervisionada será utilizado k-means e para classificação supervisionada será usado ridge regression.

2 Dataset

O dataset Mall Customers contém dados de consumidores de shopping, este dataset contém as seguintes colunas:

- CustomerID: É o ID único de um consumidor.
- Gender: Gênero do consumidor.
- Age: Idade do consumidor.
- Annual Income(k$): É a renda anual do consumidor.
- Spending Score: É a nota (1 a 100) dada a um consumidor pelo shopping baseado no dinheiro gasto e no comportamento do consumidor.

3 Clustering

Nessa seção foi realizado clustering utilizando k-means[2].

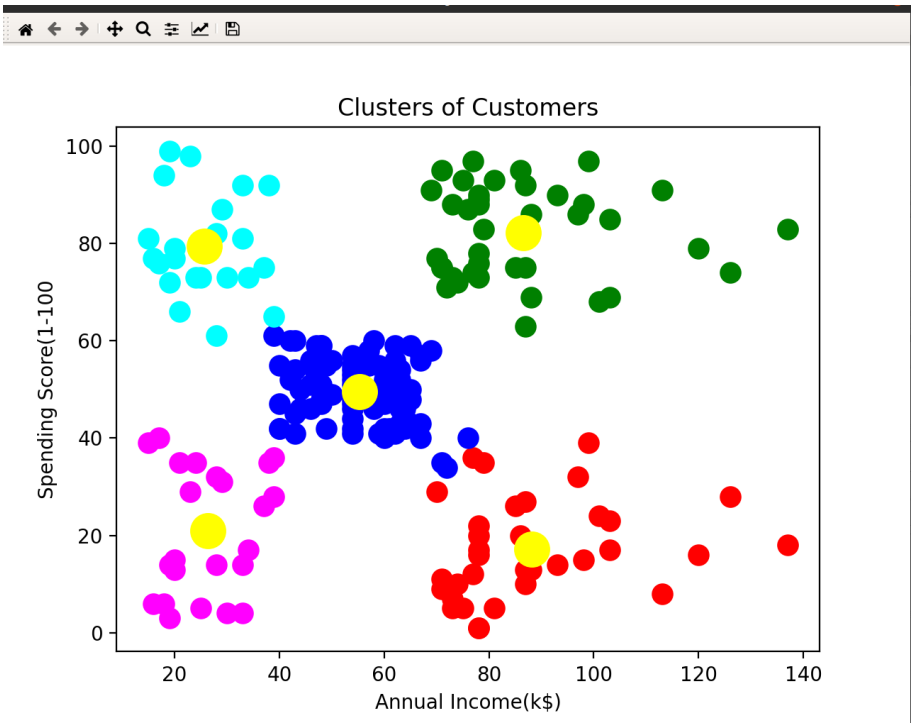

Figura 1: Dados de Customer Mall separados em 5 clusters.

É possivel identificar no cluster rosa está o grupo de pessoas com baixo renda anual e baixa nota de consumo no shopping. É algo esperado pois pessoas com baixo renda anual vão ter uma nota baixa em relação a consumo. Pode-se dizer que estas são as pessoas sensatas que sabem como gastar seu dinheiro. Este grupo não é de muita importância para o shopping.[3]

O azul claro pode-se identificar o grupo de pessoas que tem baixa renda porém tem uma nota de consumo alta. Estas são as pessoas que compram coisas mesmo tendo uma baixa renda.[3]

No grupo azul escuro podemos ver que estas pessoas tem uma renda anual mediana e uma nota de consumo média.[3]

No grupo verde é possível ver pessoas com alta renda anual e nota alta de consumo no shopping. Esse é o caso ideal para os shoppings, este grupo de pessoas são aonde os shoppings mais tem lucro.[3]

No grupo vermelho é possível identificar as pessoas que tem alta renda anual porém a nota de consumo deles é bem baixa. Estas pessoas não tem motivações para consumir no shopping. Para aumentar o lucro do shopping este grupo deve ser o maior foco de campanha para o aumento destes afim que a nota de consumo deles chegue próximo ao grupo verde.[3]

4 classificação supervisionada

Na classificação supervisionada será utilizado ridge-regression para fazer a predição. [4]

Para este cenário, alterei a coluna Male e Female dos dados para 0 e 1 pois o algoritmo estava esperando número e não string.

Para uma nova linha de row = [19,1,52,23] tivemos o resultado de "Predicted: 40.903"e quando usamos um valor de linha de [19,1,52,23] temos um aumento para "Predicted: 67.361"E se gerarmos um cenário mais difícil de acontecer com os valores row = [100,0,100,100] podemos esperar uma probabilidade bem baixa e o resultado não é diferente: "Predicted: 15.087"

5 método de redução de dimensionalidade

Para o método de redução utilizamos o Univariate feature selection, conforme a documentação, Univariate feature selection funciona por selecionar as melhores qualidades baseado em univariate statistical tests. Pode ser visto como um passpo de pré processamento a um estimador. Para essa seção foi utilizado o Univariate feature selection do scikit-learn. Aplicando no mesmo dataset com k=2 e alternando para 2 clusters utilizando o k-means da primeira seção, temos a seguinte distribuição.[5]

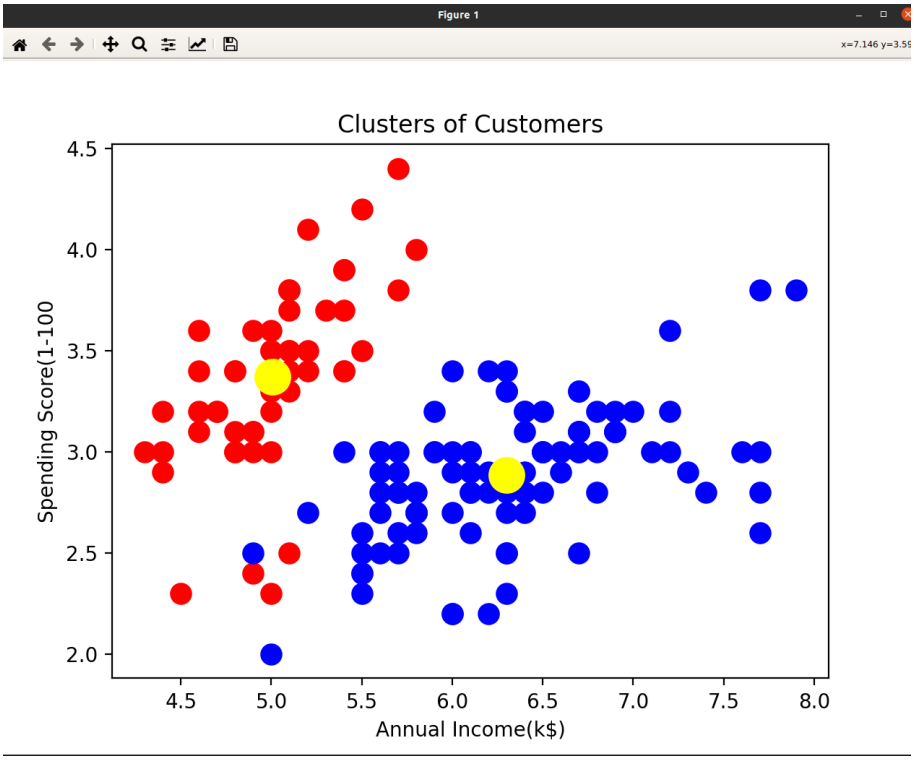

Figura 2: Aplicando redução de dimensionalidade.

6 Conclusão

Este trabalho pode ser baixado e executado em https://github.com/guilfreis/IA-mall-customer

Referências

[1] shwetabh123. Mall customers data, 2017.

[2] Samet Girgin. K-means clustering model in 6 steps with python, 2019.

[3] Shubhankar Rawat. Mall customers segmentation — using machine learning, 2019.

[4] Jason Brownlee. How to develop ridge regression models in python, 2020.

[5] Scikit learn. 1.13. feature selection.