INSTITUTO TECNOLÓGICO DE AERONÁUTICA

CMC-13 - Introdução à Ciência de Dados

Projeto Aprendizado de Máquina

Bruna Belarmino Silva

Prof. Paulo André Castro

1. Classificador baseado em árvore de decisão

1.1. Preparação dos dados

Os dados foram tratados de forma a deixá-los prontos para serem apresentados para a árvore de decisão implementada.

A tabela de filmes apresenta 3883 filmes. Importando os dados, e tratando para pular linhas com erros, foi obtido uma tabela com 3882 filmes, o que é considerado muito bom, pois só um filme foi pulado.

A tabela de usuários era composta por UserID, Gender, Occupation, Zip-code, name e birthday. Essa tabela é formada por 6040 usuários. O atributo "zip-code" foi eliminado, uma vez que a base de teste foi feita para ser testada por uma pessoa brasileira. Além disso, essa eliminação permitia uma análise mais simplificada e, como a base de teste era pequena, não houve grandes impactos na análise. O atributo "name" foi desconsiderado, uma vez que não faz sentido o nome influenciar na classificação que a pessoa dá para o filme. O atributo "birthday" foi utilizado para fornecer a idade e, assim, separar em grupos de faixa etária. Durante esse tratamento, foi observado que alguns meses estavam com problemas, uma vez que indica 0, o que não faz sentido. Para tratar esse dado inconsistente, foi atribuído o mês 1 e posteriormente a idade foi calculada. Note que uma transformação possível seria apenas considerar o ano e subtraí-lo de 2022. No entanto, para não dar possíveis divergências de ano sem considerar o mês, para casos limites de faixa etária, foi escolhido manipular os dados assim. Em seguida, a coluna birthday foi eliminada, uma vez que o dia e mês em que a pessoa nasceu não são fatores relevantes para essa análise.

A tabela de classificação é composta por UserID, MovieID, Rating e Timestamp. O Timestamp foi eliminado, uma vez que esse código associado à data da postagem de avaliação não traz informações relevantes a ponto de serem consideradas para a análise.

Após verificar que não havia linha com dados faltantes, as tabelas foram relacionadas de forma a ser apresentada para a árvore de decisão. Para isso, foi utilizado o MovieID e o UserID. Depois, o título do filme foi eliminado, uma vez que existe uma relação de 1 para 1 entre Title e MovieID e, para a árvore implementada, assim como bibliotecas que implementam árvores de decisão, é necessário fazer essa transformação de atributos categóricos em dados numéricos. Igualmente, o 'Genre' e 'Gender' também foram transformados em números. Finalmente, os atributos foram divididos entre previsores e classe. Os previsores foram: MovieID, age_group, Genre, Gender e Occupation. A classe é o rating.

1.2. Resultados Obtidos

Ao implementar a árvore de decisão, foi obtido que a raiz da árvore é o MovieID. Para nível de teste, foi utilizada uma tabela com minhas informações de gênero, ocupação, grupo de idade, filme, gênero e a classificação dada para o filme por mim. A acurácia para esse teste foi de 0.3. É importante ressaltar que esse valor não é significativo, pois o projeto não utilizou uma boa base de treinamento.

Foi obtida a seguinte matriz de confusão:

$$\text{Matriz} = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 4 & 1 \end{matrix}$$

O erro quadrático médio obtido foi de 1.3.

Imagem: Estrutura da árvore de decisão

```
Out[67]: {'MovieID': {1: {'Occupation': {0: {'age_group': {1.0: {'Gender': {0: {'Genres': {145: 5.0}},
                                                              1: {'Genres': {145: 4.0}}}},
                                                    18.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 5.0}}}},
                                                    25.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 4.0}}}},
                                                    35.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 5.0}}}},
                                                    45.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 5.0}}}},
                                                    55.0: {'Gender': {0: {'Genres': {145: 2.0}}, 1: {'Genres': {145: 4.0}}}},
                                                    56.0: {'Gender': {0: {'Genres': {145: 4.0}},
                                                      1: {'Genres': {145: 4.0}}}}}},
                                                  1: {'age_group': {1.0: {'Gender': {0: {'Genres': {145: 4.0}},
                                                      1: {'Genres': {145: 5.0}}}},
                                                    18.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 4.0}}}},
                                                    25.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 5.0}}}},
                                                    35.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 4.0}}}},
                                                    45.0: {'Gender': {0: {'Genres': {145: 4.0}}, 1: {'Genres': {145: 4.0}}}},
                                                    55.0: {'Gender': {0: {'Genres': {145: 5.0}}, 1: {'Genres': {145: 3.0}}}},
                                                    56.0: {'Gender': {0: {'Genres': {145: 3.0}},
                                                      1: {'Genres': {145: 3.0}}}}}},
```

O classificador a Priori considera apenas as médias de avaliação de cada filme. Para os filmes avaliados foram obtidas as seguintes médias:

Filme	Média	Filme	Média
Toy Story (1995)	4	Jurassic Park (1993)	4
Once Upon a Time… When We	4	Godzilla 2000 (Gojira ni-sen	3

Were Colored (1995)		mireniamu) (1999)	
Lion King, The (1994)	4	Blade Runner (1982)	4
Walking Dead, The (1995)	3	Titanic(1953)	3
Browning Version, The (1994)	4	Modern Times (1936)	4

A acurácia do classificador a priori foi 0.3. Note que é o mesmo valor para o classificador baseado em árvore de decisão. Esse fato contribui para invalidar o valor da acurácia desse classificador (árvore de decisão), uma vez que os dados de teste não são adequados.

Foi obtida a seguinte matriz de confusão:

$$\text{Matriz} = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{matrix}$$

O erro quadrático médio foi de 1.6. Como esse valor foi maior em relação ao do classificador com árvore de decisão, então, para esse conjunto de teste, o primeiro classificador (baseado em árvore) funcionou melhor, uma vez que houve menor flutuação em torno dos valores corretos.

Comparando-se em termos de estatística Kappa, o coeficiente do classificador a priori foi -0.06 e o do classificador baseado em árvore de decisão foi 0.167. Isso indica que o classificador baseado em árvore tem maior concordância em relação ao a priori.

2. Conclusões

Primeiramente, conclui-se que como o MovieID é a raiz, então o gênero não precisava ter sido implementado na árvore de decisão pois, determinando-se o filme, o gênero já está determinado. Isso poderia ter sido previsto e evitado se fosse calculado e analisado o ganho de informação de cada conjunto.

No que tange aos resultados, foi possível concluir que, embora a acurácia não tenha fornecido informações tão razoáveis para a comparação dos classificadores em termos de uma pequena base de teste, outros métodos utilizados indicaram que o classificador baseado em árvore de decisão é melhor em relação ao classificador a Priori.

A respeito do desenvolvimento do trabalho, notou-se que a implementação da árvore de decisão possui uma alta complexidade, o que torna evidente a necessidade de utilização de bibliotecas que realizam esse procedimento.

3. Descrição da implementação

O projeto foi desenvolvido utilizando Python, desenvolvido no Jupyter Notebook como IDE. A biblioteca plotly foi instalada apenas para facilitar a análise dos dados.