

Introdução

O presente documento chamado Projeto IV, tem como o objetivo apresentar os resultados obtidos durante a implementação e estudos realizados nos tópicos de Detecção de Anomalia e Sistemas de Recomendação que faz parte do conteúdo da disciplina "Aprendizado de Máquinas". A primeira parte do documento faz referência à Detecção de Anomalia em servidores computacionais usando o modelo Gaussiano multivariado. Para realizar o anteriormente mencionado empregamos o arquivo *dado1.mat* que contém um conjunto de servidores compilado na matriz Y , cada um dos exemplos possui o tempo específico levado para chegar ao destino ou também chamado latência posicionado na primeira coluna da matriz Y e a quantidade de dados transferidos de um lugar para outro ou a taxa de transferência estão na segunda coluna da matriz Y .

Já que uma anomalia se caracteriza por uma baixa probabilidade, menor que um threshold ϵ de aquele dado pertencer à distribuição Gaussiana obtida usando como medida de avaliação o $F_1 - score$, empregamos os dados *Xval* e *yval* do arquivo *dado2.mat*, pois ele tem mais servidores e mais atributos para calcular os ϵ , $F_1 - score$ e o número de anomalias.

A segunda parte do projeto é todo o relacionado aos sistemas de recomendação, donde usamos o arquivo *dado3.mat* que contém notas de 1 a 5 dadas por usuários para filmes. A matriz Y armazena na linha i e coluna j a nota dada pelo usuário j para o filme i . Ademais, contém a matriz R onde temos $R(i, j) = 1$ se o usuário j deu alguma nota para o filme i e 0 caso contrário. Com este dados implementamos o algoritmo de filtragem colaborativa considerando a função de custo sem regularização e use gradiente conjugado para minimizá-la. Por ultimo, listamos os 10 filmes com notas médias mais altas, mostrando o nome e a nota média do respectivo filme, para tudo isto utilizamos o arquivo *dado4.txt* que contém o nome correspondente a cada linha na matriz Y .

Parte I

Detecção de Anomalia

Algoritmos de detecção de anomalias podem ser usados para monitorar o acesso aos dados e sinalizar variações da norma, por exemplo, identificando quando as pessoas acessam dados em pontos de tempo que são anômalos em comparação com seu histórico de acesso anterior e o histórico de acesso de outras pessoas com funções semelhantes na mesma organização. Graças à detecção precoce de falhas, as falhas fatais são evitadas e as tarefas de manutenção podem ser agendadas quando forem mais econômicas. As tecnologias de manutenção preditiva são essenciais para estender a vida útil do equipamento, reduzindo os custos de manutenção e aumentando a exploração dos ativos.

As anomalias podem ser definidas como padrões ou pontos de dados que não estão de acordo com uma noção bem definida de comportamento normal. Em contraste com a remoção de ruído e acomodação de ruído, onde as anomalias são vistas como um obstáculo para a análise de dados, a detecção de anomalias permite uma análise de dados interessante com base na identificação das anomalias.

Para detectar anomalias quando o espaço do recurso original foi reduzido perdendo alguns dados, mas esperando que se retenha a variação mais importante, ao lidar com uma ou duas variáveis, a visualização de dados geralmente pode ser um bom ponto de partida. No entanto, ao dimensionar isso para dados de alta dimensão, essa abordagem se torna cada vez mais difícil.

As coleções de pontos de dados, geralmente têm uma certa distribuição (por exemplo, uma distribuição Gaussiana). Para detectar anomalias de uma forma mais quantitativa, primeiro calculamos a distribuição de probabilidade $p(x)$ dos pontos de dados. Então, quando um novo exemplo x aparece, comparamos $p(x)$ com um limite ϵ . Se $p(x) < \epsilon$ é considerado uma anomalia. Isso ocorre porque os exemplos normais tendem a ter uma grande $p(x)$, enquanto os exemplos anômalos tendem a ter uma pequena $p(x)$.

Na teoria dos dados, presume-se que cada $x \in \mathbb{R}$ segue uma distribuição Gaussiana Normal própria com média μ e variância σ^2 :

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

com a probabilidade que indica que $p(x)$ é parametrizado por μ e σ^2 :

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Os parâmetros μ e σ^2 são obtidos do conjunto de treinamento como,

$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)^2$$

Como já foi mencionado, empregamos o arquivo *dado1.mat* que contém um conjunto de exemplo de servidores copilados na matriz Y , na primeira coluna contém sua respectiva latência e na segunda coluna tem a taxa de transferência. A latência informa o tempo transcorrido para uma informação ou pacote percorrer o caminho de ida e volta de origem para destino, é medida em milissegundos (ms), ela é importante porque algumas aplicações são muito sensíveis à latência. Já a taxa de transferência é o número médio de bits, caracteres ou blocos convertidos ou processados por unidade de tempo que passam entre equipamentos num sistema de transmissão de dados. Comumente é medido em bits por segundo (b/s). Estas medidas são muito úteis para determinar se um servidor está se comportando normalmente. Porém, uma grande maioria de servidores suministrados no arquivo são "normais" e alguns poucos "anômalos", como pode ser observado na figura 1.

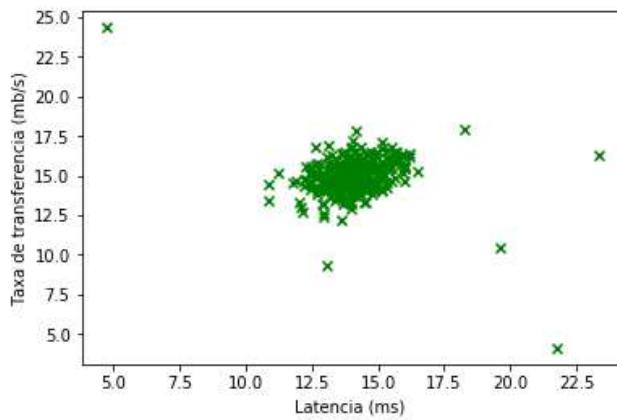


Figura 1: Plot Latência vs Taxa de Transferência

Agora, implementamos uma rotina ajustada com uma distribuição Gaussiana normal multivariada, esta distribuição é uma ferramenta muito poderosa para descobrir anomalias ou outliers porque esse algoritmo também leva em conta como as variáveis mudam com outras variáveis no conjunto de dados, o que muitos outros algoritmos não fazem. O vetor aleatório $x \in \mathbb{R}^n$ é normal com o vetor de média $\mu \in \mathbb{R}^n$ e a matriz de covariância $\Sigma \in \mathbb{R}^{n \times n}$,

$$x \sim \mathcal{N}(\mu, \Sigma)$$

com a probabilidade que indica que $p(x)$ é parametrizado por μ e Σ :

$$p(x; \mu, \Sigma) = \frac{1}{2\pi^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

em que $|\Sigma|$ é o determinante de Σ .

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

Sendo que Σ é positiva definida, sua inversa também o é, e a forma quadrática $(x - \mu)^\top \Sigma^{-1}(x - \mu)$ é sempre positiva para qualquer vetor x não nulo. Para o caso dos dados da figura 1 temos $n = 2$, logo esta equação define uma elipse no plano (x_1, x_2) chamada elipse de concentração. A excentricidade e inclinação dos eixos da elipse dependem dos parâmetros da matriz de covariância. Nessas figuras, pode-se ver o significado da correlação das variáveis, mas o fato de suas variâncias serem diferentes apenas modifica a excentricidade das elipses de concentração, mas não a relação entre as variáveis.

Neste caso particular, a matriz de covariância é uma matriz diagonal pois a elipse do gráfico 1 não tem rotação, isso significa que não existe correlação entre a latência e a taxa de transferência.

Posto que, todos os modelos gerarão alguns falsos negativos, alguns falsos positivos e possivelmente ambos e para responder o quão preciso é um modelo é essencial otimizar as métricas de desempenho mais úteis e conhecer a precisão, a recuperação e as pontuações $F_1 - score$, as quais são medidas gerais da precisão de um modelo que combina precisão e recall. Ou seja, uma boa pontuação de F_1 significa que tem poucos falsos positivos e poucos falsos negativos, portanto, está identificando corretamente ameaças reais e não é incomodado por alarmes falsos. Uma pontuação

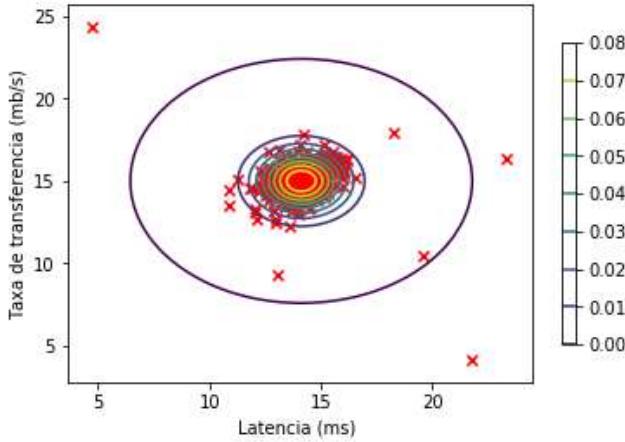


Figura 2: Gráfica das elipses de concentração dos dados Y_1 e Y_2

$F_1 - score$ é considerada perfeita quando é 1, enquanto o modelo é uma falha total quando é 0.

Do algoritmo empregado obtemos que, o melhor epsilon ϵ obtido por validação é $8.990852779269495e - 05$ e o melhor $F_1 - score$ na validação é 0.8750000000000001 . o que significa que possui poucos falsos positivos e falsos negativos.

Na literatura da ciência de dados, podem existir três tipos de anomalias, a compreensão desses tipos pode afetar significativamente o modo como você trata as anormalidades mas o que sim é seguro as anomalias não podem formar um grupo denso, pois os estimadores disponíveis presumem que os outliers estão em regiões de baixa densidade. Os tipos de anomalias são,

- *Anomalias globais* são o tipo mais comum de anomalias e correspondem aos pontos de dados que se desviam muito do resto dos pontos de dados, como se pode observar na figura 3. O principal desafio é descobrir a quantidade exata de desvio que leva a uma anomalia potencial.
- *Anomalias contextuais* são aquelas em que o desvio que leva à anomalia depende da informação contextual. Esses contextos são regidos por atributos contextuais e atributos comportamentais. A figura 4 mostra os dados de uma série temporal para um determinado período de tempo. Os valores não ficaram fora dos limites globais normais, mas na verdade existem pontos anormais (laranja) em comparação com a sazonalidade.

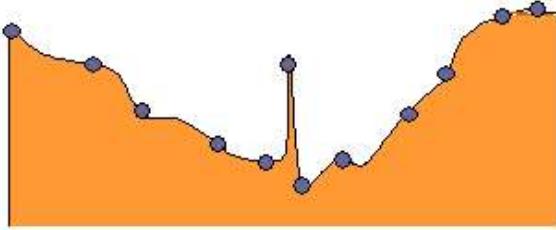


Figura 3: Anomalias Glorais

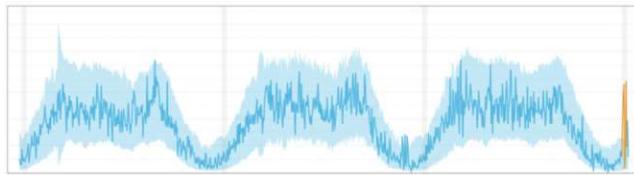


Figura 4: Anomalias Contextuais

- *Anomalias coletivas* são que os pontos de dados incluídos na formação da coleção, mas elas podem não ser anomalias quando considerados individualmente. Elas são interessantes porque não é preciso olhar para pontos de dados individuais, se não analisar seu comportamento coletivo. Na figura 5, os pontos de dados marcados em verde formaram coletivamente uma região que se desvia substancialmente do resto dos pontos de dados.

Logo, segundo o anterior mencionado e a figura 6 o problema tem anomalias globais pois existem pontos que estão muito distantes dos outros, seis pontos para ser mais preciso. Além disso, na figura 7 podemos observar a maior elipse de concentração que circula as anomalias com $p(x) < 8.990852779269495e - 05 \approx 0.000090$.

Visto que as técnicas de detecção de outliers podem ser realizadas numericamente ou graficamente, é por isto que podemos utilizar a validação cruzada para obter os valores ótimos de ϵ e $F_1-score$; para isto executamos o sistema desenvolvido sobre o arquivo *dado2.mat* porque este possui 11 atributos (*Xval* e *yval*). Assim, o melhor ϵ obtido por validação é $1.3772288907613575e - 18$ e o melhor $F_1-score$ na validação é 0.6153846153846154. O total de anomalias que possui o conjunto original é de 117 e a percentagem de anomalias é 0.1%.

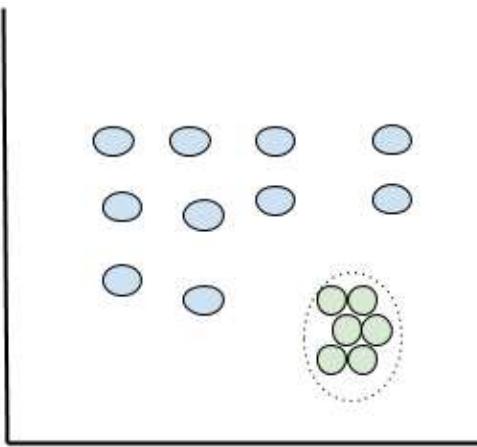


Figura 5: Anomalias Coletivas

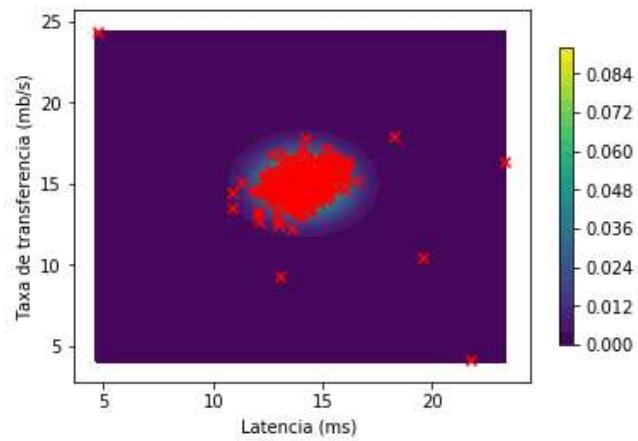


Figura 6: Plot das Anomalias

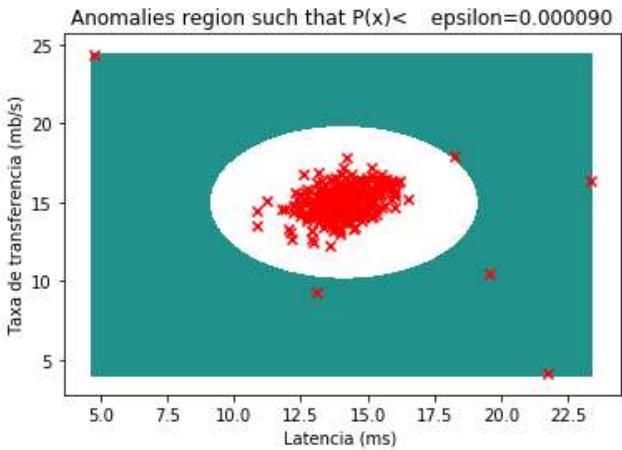


Figura 7: Plot das Anomalias com $p(x) \leq 0.000090$

Parte II

Sistemas de Recomendação

Na computação, os sistemas de recomendação estudam as preferências e gostos dos usuários para sugerir itens viáveis e interessantes. Um sistema de recomendação auxilia ao usuário a filtrar informações relevantes com base em uma série de critérios ou objetivos, sejam eles preferências, gostos ou necessidades, que constituem o perfil personalizado de um determinado usuário. Ou seja, é um sistema inteligente que oferece aos usuários sugestões personalizadas sobre um determinado conteúdo.

Ultimamente, esses sistemas começaram a usar algoritmos de aprendizado de máquina para prever o processo e encontrar itens mais adequados. Com base nas informações recebidas dos sistemas de recomendação, os algoritmos mudam. Os algoritmos de aprendizado de máquina de sistemas de recomendação são geralmente divididos em uma das categorias: filtragem colaborativa e baseada em conteúdo.

A filtragem colaborativa é uma técnica que ajuda a obter melhores resultados de sistemas de recomendação. Possui um funcionamento regido por algoritmos matemáticos que classificam as informações, estudam-nas e geram sugestões ajustadas às necessidades do usuário. É dizer, para cada usuário é indicado um item baseado em um outro usuário com um “gosto similar”.

Uma vantagem da abordagem de filtragem colaborativa é que ela não é baseada

em conteúdo analisável por máquina e, portanto, é capaz de recomendar com precisão elementos complexos, como filmes, sem precisar "entender" o próprio elemento.

Para iniciar o algoritmo definimos o número de usuários n_u , o número de filmes n_m , a matriz $R(i, j)$ onde $R(i, j) = 1$ se o usuário j deu alguma nota para o filme i e 0 caso contrário; $\theta^{(j)}$ como o vetor de parâmetros para o usuário j e o $x^{(j)}$ como o vetor de atributos para o filme i , então, para o usuário j e filme i , a nota predita é dada por

$$(\theta^{(j)})^\top (x^{(i)})$$

Para aprender os parâmetros $\theta^{(j)}$ resolvemos o problema de otimização dado por

$$\arg \min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} ((\theta^{(j)})^\top (x^{(i)}) - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

Uma vez obtidos os $\theta^{(j)}$, estimamos os parâmetros $x^{(i)}$, os quais devem ser minimizados para estimar os $\theta^{(j)}$. Para isso, resolvemos o problema de otimização

$$\arg \min_{x^{(1)}, \dots, x^{(n_m)}} \frac{1}{2} \sum_{j=1}^{n_m} \sum_{i:r(i,j)=1} ((\theta^{(j)})^\top (x^{(i)}) - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n_m} \sum_{k=1}^n (x_k^{(j)})^2$$

Note que os valores de $x^{(j)}$ dependem dos valores de $\theta^{(j)}$ e vice-versa. Logo, o problema anterior pode ser juntado em um só:

$$\arg \min_{\substack{x^{(1)}, \dots, x^{(n_m)}, \\ \theta^{(1)}, \dots, \theta^{(n_u)}}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} ((\theta^{(j)})^\top (x^{(i)}) - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n_m} \sum_{k=1}^n (x_k^{(j)})^2 \quad (1)$$

Os $\theta^{(1)}, \dots, \theta^{(n_u)}$ e $x^{(1)}, \dots, x^{(n_m)}$ são inicializados com valores aleatórios pequenos e resolver o problema de minimização dado em (1) pode ser usado qualquer método de otimização, por exemplo o gradiente conjugado. Ademais, para simplificar os cálculos os dados podem ser vetorizados definindo as matrizes X com $(x^{(i)})^\top$ em cada linha e Θ com $(\theta^{(j)})^\top$ em cada linha. A matriz de predição pode ser representado por $X\Theta^\top$, chamado fatoração de matriz de posto baixo pelo fato de ter posto baixo.

O conceito de similaridade é um complemento a distância. Quando falamos de semelhança, ao invés de descobrir o quanto longe dois pontos estão (distância euclidiana), nosso objetivo é descobrir o “quanto perto” os pontos estão. Dos filmes i e j são relacionados ou similares se

$$\|x^{(i)} - x^{(j)}\|$$

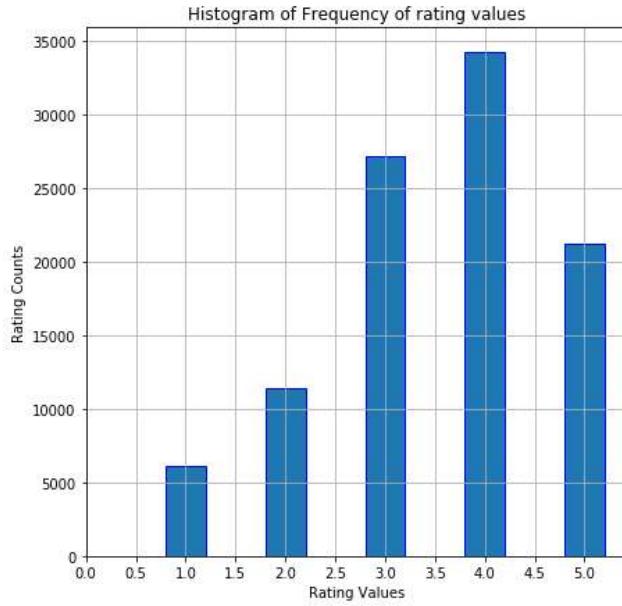


Figura 8: Histograma de frequêcia dos valores de classificação dos atributos do arquivo *dado3.mat*

for pequeno, por isso a similaridade também é chamada de “anti-distância”.

Na implementação do algoritmo de filtragem colaborativa foi usada a base de dados *dado3.mat* que contem filmes com sua respetiva matriz R como foi definida inicialmente e a matriz Y com as notas dos usuários, em que y_{ij} tem a nota dada pelo usuário j para o filme i . Considerando o problema de otimização dado em (1) e usando gradiente conjugado para minimizá-lo com 400 iterações e $\lambda = 0$, este algoritmo implementa aprendeu a matriz X , em que cada linha contem o vetor de atributos $x^{(i)}$ do i -ésimo filme, e a matriz Θ que em cada linha guarda o vetor de parâmetros $\theta^{(j)}$ para o j -ésimo usuário. A figura 8 é o histograma de frequêcia das notas dadas pelos usuários para todos os filmes, ignorando aqueles que não tem nota, isso é, aqueles que tem $R_{ij} = 0$.

Com base as notas preditas pelo algoritmo, os 10 filmes com notas médias mais altas estão na figura 9. Observa-se que aproximando ao inteiro mais próximo, o promédio é acertado para os filmes melhor qualificados.

É para destacar que o total de filmes não avaliados para todos os usuários é de 1486126, é por isto que, para eliminar o caso de algum usuário não classificar o

	Film	Predicted Mean	Real mean
0	1189 Prefontaine (1997)	5.001232	5.0
1	1293 Star Kid (1997)	5.001042	5.0
2	1467 Saint of Fort Washington, The (1993)	5.000869	5.0
3	1599 Someone Else's America (1995)	5.000829	5.0
4	1122 They Made Me a Criminal (1939)	5.000314	5.0
5	1653 Entertaining Angels: The Dorothy Day Stor...	5.000226	5.0
6	1201 Marlene Dietrich: Shadow and Light (1996)	4.999238	5.0
7	1500 Santa with Muscles (1996)	4.998954	5.0
8	814 Great Day in Harlem, A (1994)	4.998869	5.0
9	1536 Aiqing wansui (1994)	4.998451	5.0

Figura 9: Os 10 filmes com as médias mais altas

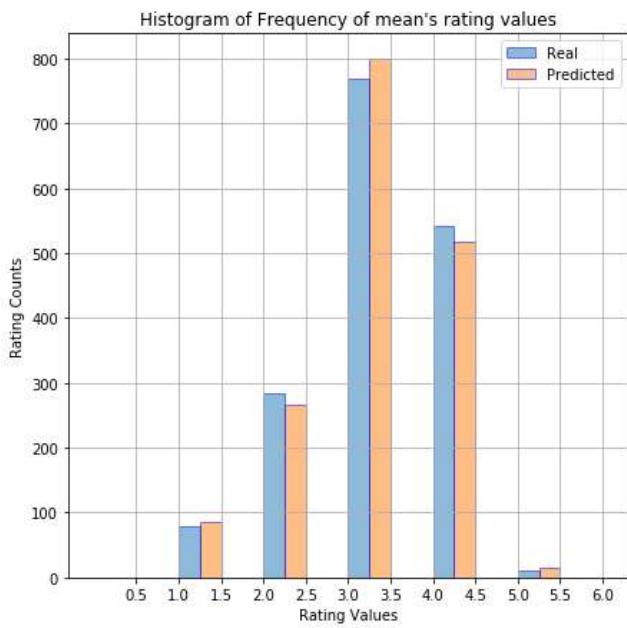


Figura 10: Histograma de frequênciade classificação de médias dos atributos do arquivo *dado4.mat*

filme normalizamos as classificações. Outra consequência do grande numero de não avaliados é que se analisarmos cada usuário, podem ter alguns em que a nota predita para algum filme é superior a 5, isso para que a média predita seja cercana da verdadeira média.

Já na figura 10 está o histograma de frequênciadas notas médias para cada filme, comparando a média predita e a média real. Pode-se observar que o comportamento da quantidade de filmes para cada nota é similar. Agora, gostaríamos de comparar a média real e predita para cada filme, temos que norma de erro exato é de 0.8443810314799114, em que se refere a erro exato como o diferença das medias sem aproximar. Finalmente, se aproximamos ao inteiro mais próximo, obtemos que a porcentagem da nota media para cada filmes que foram bem previstos é 97.08680142687277%.

Anexos

A seguinte tabela contem as anomalias para o arquivo *dado2.mat*, em que é denominado outlier se $p(x) \leq 1.3772288907613575e - 18$.

	$p(x)$
Index in X	
9	2.997653e-19
20	7.349666e-21
21	1.292807e-19
30	1.619280e-21
39	1.005373e-18
56	4.731257e-19
62	9.315961e-19
63	1.155380e-18
69	1.051353e-18
70	1.208939e-18
77	8.409412e-19
79	9.744901e-23
86	7.894546e-19
103	3.605101e-19
130	5.966535e-20
147	1.282068e-18
154	1.077505e-18
166	1.097231e-18
175	1.310352e-20
176	1.219238e-18
198	1.577046e-19
209	2.334045e-19
212	5.068854e-20
218	4.614413e-19
222	2.663336e-20
227	8.324935e-20
229	1.257346e-18
233	8.144030e-19
244	1.129762e-18
262	1.611848e-20
266	4.474960e-20
271	5.339471e-19
276	1.780480e-19

	$p(x)$
Index in X	
284	1.334061e-18
285	2.829521e-19
288	4.890279e-19
289	8.312446e-19
290	6.728567e-19
297	1.024621e-18
303	5.007886e-21
307	4.780885e-19
308	8.907541e-19
320	2.263329e-19
324	1.207666e-18
338	1.824867e-19
341	1.220420e-19
342	5.685207e-19
344	2.140617e-19
350	9.570968e-19
351	6.140376e-19
353	1.263210e-19
365	3.821437e-19
369	1.181639e-18
371	4.200767e-19
378	8.356898e-19
398	2.069280e-19
407	6.610194e-20
420	9.315251e-21
421	1.659979e-26
424	1.001499e-19
429	8.074607e-19
438	1.302621e-18
452	5.148058e-19
455	1.215032e-19
456	4.164854e-22
462	7.130409e-20
478	3.174032e-23
497	9.650525e-19
518	4.613973e-19
527	3.453240e-19

Index in X	p(x)
530	2.618313e-19
539	7.085023e-19
541	2.404567e-19
551	4.010248e-20
574	5.494259e-19
583	1.146393e-18
587	6.049171e-20
602	7.871508e-19
613	1.188672e-18
614	8.949930e-19
628	3.438809e-19
648	4.286490e-21
674	7.793822e-22
678	3.185701e-20
682	6.722133e-20
685	2.592348e-22
700	1.008424e-19
702	1.831491e-23
705	5.067007e-19
713	7.438547e-19
721	1.722640e-20
741	2.881505e-19
750	1.205391e-18
757	1.132466e-18
758	1.091835e-18
787	7.670506e-19
831	7.445229e-21
834	1.214554e-18
836	5.016248e-20
839	1.036009e-18
846	3.875261e-19
870	7.140305e-19
885	1.597437e-20
887	1.009863e-18
890	3.291406e-19
901	3.524364e-20
911	4.562561e-20

	$p(x)$
Index in X	
930	1.878151e-19
939	1.729905e-19
940	9.374402e-19
943	2.992625e-19
951	8.777746e-19
952	3.145989e-20
970	7.555243e-19
975	6.207761e-19
992	2.191264e-19
996	1.532452e-19

Bibliografia

- [1] Juliano José da Silva Santos, Tânia Lucia Graf de Miranda, and Rafael Geha Serta. Modelagem da dispersão de poluentes de origem móvel (veicular) em curitiba e a ocorrência de bromélias em cabos da rede elétrica. *Blucher Engineering Proceedings*, 1(2):189–204, 2014.
- [2] Aplicación De La Transformada Wavelet En, Y Dos Dimensiones Para El Análisis, Compresión De, and Lorena Paola Pazmiño Altamirano. Carrera de ingeniería en electrónica y telecomunicaciones.
- [3] T Hastie, R Tibshirani, and J Friedman. Springer series in statistics the elements of statistical learning data mining, inference. Technical report, and Prediction (Tech. Rep.). Retrieved from <https://web.stanford.edu/> . . . , 2001.