Universidade Federal de Ouro Preto

Mateus Oliveira dos Santos - 11.2.8093

Exercício Extra

Trabalho Extra apresentado a disciplina de Avaliação e Desempenho de Sistemas Computacionais do curso de Engenharia de Computação do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto.

Professor Alexandre Magno de Sousa.

João Monlevade - MG
Fevereiro 2016

1 Enunciado

1. Um servidor de banco de dados possui uma CPU e dois discos, o monitor de desempenho do SGBD, que realiza medições em nível de transações, gerou um log de atividades dos recursos do sistema para cada transação que ocorreram em um intervalo de 2 minutos e meio. Ao todo 200 transações foram registradas nesse período de tempo conforme os dados da planilha anexa "DBMS-Performance-Monitor-Log.xls" (tempo de CPU e número de I/Os para cada disco, além do ID para cada transação). Para realizar a caracterização do workload do servidor, faça o que se pede:

(a) apresente uma tabela com informações com estatísticas básicas para cada feature registrada, tais como: média, variância, desvio padrão, coeficiente de variação, valor total (soma), valores mínimo e máximo, range (faixa de valores abrangido, e.g. máximo - mínimo), 1o quartil, 2o quartil e 3o quartil.

(b) apresente gráficos para cada feature tais como: histograma de único parâmetro, *Cumulative Distribution Function* (CDF) e Box Plot.

(c) realize a Análise do Componente Principal (PCA) seguindo todos os passos apresentados em sala de aula, não se esqueça de construir os gráficos.

(d) construa uma tabela de correlação conforme a Tabela 1, em que $R_{(i,j)}$ representa a correlação da feature i com o principal fator j calculado pelo PCA de acordo com a equação

$$R_{(i,j)} = \frac{\frac{1}{n-1}\sum_k^n (i_k - \hat{i})(j_k - \hat{j})}{S_i S_j} \tag{1.1}$$

onde n representa o número total de componentes registrados, $\hat{i}$ e $\hat{j}$ são médias, S_i e S_j são o desvio padrão, e i_k e j_k representam os dados de cada componente da *feature* i e do principal fator j.

A partir dessa tabela, elabore um gráfico do tipo plano cartesiano onde cada *feature* (Tempo de CPU, # de I/Os do disco 1 e 2) sejam representados como pontos (x, y) onde x é o valor de correlação com o Principal Fator 1 e y é o valor de correlação com o Principal Fator 2. Esse gráfico mostrará a relação dos dados medidos com os principais fatores, geralmente é mais utilizado para dados com alta dimensionalidade, $n > 3$, isto é, quanto maior os valores de correlação com os principais fatores y_1 e y_2, maior é a influência desses fatores na componente de dados.

(e) mostre os resultados, faça observações e apresente conclusões

Tabela 1 – Tabela de análise PCA

Programa	Principal Fator 1 (y_1)	Principal Fator 2 (y_2)
Tempo de CPU	$R_{(cpu,y1)}$	$R_{(cpu,y_2)}$
# I/Os Disco 1	$R_{(I/OsDisco1,y_1)}$	$R_{(I/OsDisco1,y_2)}$
# I/Os Disco 2	$R_{(I/OsDisco2,y_1)}$	$R_{(I/OsDisco2,y_2)}$

Observação: para criação dos gráficos CDF e Box Plot, que não foram apresentados na disciplina, realize uma pesquisa em estatística para ter conhecimento de como se constrói esse tipo de gráfico.

2 Introdução

O trabalho apresenta a resolução do problema proposto na seção de enunciado 1. A seção 3 apresenta as observações e conclusões obtidas a partir da análise dos dados resultantes das seções de 4 a 7. As seções seguem a ordem da resolução do problema proposto, exceto pela questão **e** que será a primeira a ser apresentada.

3 Análise dos resultados - Questão E

Apresentaremos as conclusões e observações de acordo com a sequencia de resolução das questões.

3.1 Questão A

As tabelas 3 e 2 apresentam os dados estatísticos do problema e os dados reais parciais, para acesso aos dados completos acesse o repositório do git citado em 4.1. Como esperado os dados normalizados possuem média zero e desvio padrão um. Além dos somatórios igual a zero das componentes normalizadas e dos fatores do PCA. As observações dos gráficos explicaram melhor o que cada resultado significa na seção 3.2.

3.2 Questão B

Analisando os histogramas na seção 5.1 podemos observar que as concentrações de cargas que tomam maior de tempo de cpu se encontram entre 75 ms e 172ms, e uma segunda carga aparentemente média entre 412 ms e 508ms. Para o disco 1 a maioria das transações, cerca de 50, executam aproximadamente 16 IOs enquanto que para o disco 2 são 30 IOs por transação, com uma frequência igual ao disco 1. No disco um as demais transações estão mais distribuídas, enquanto que no disco 2 as transações com maior quantidade de IOs estão entre 64 e 80 IOs por transação.

Ao se observar os gráficos das CDFs 5.2 podemos fornecer um dado mais preciso, onde cerca de 60% das operações na CPU gastam aproximadamente 400ms de tempo de CPU. No disco 1 apenas 30% das transações fazem até 30 IOs , já no disco 2 aproximadamente 26% das transações fazem até 40 IOs.

Com o box plot conseguimos entender melhor as curvas das CDFs dos disco, como podemos ver no gráfico da figura 8 enquanto do disco 1 apresenta maior variabilidade entre o primeiro quartil e a mediana, o disco 2 é o oposto. Apesar dos máximos e mínimos estarem próximos os tipo de carga que cada disco recebe varia claramente em número de IOs, ou seja, o disco dois recebem operações que fazem mais IOs que o disco 1.

O box plot do CPU 7 mostra que a maior parte das operações tem variabilidade em tempo de CPU entre a mediana e o terceiro quartil.

3.3 Questão C

Realizando o PCA encontramos que o principal fator 1 é o disco 1, com cercar de 58% de representação dos dados e o principal fator 2 é a CPU, com cercar de 35% de representação dos dados. Logo, o disco 2 com aproximadamente 7%.

Observando o gráfico 9 podemos observar que, apesar de uma dispersão próximo ao valor dois do principal fator 2 (CPU), a maior parte dos dados se concentram ao longo do eixo do principal

fator 1 (Disco 1). De certa forma, era de se esperar que o principal fator 1 tinha grande chances de ser um dos disco por se tratar de um sistema de banco de dados.

3.4 Questão D

Por fim, observando o gráfico 10 onde o eixo x é o principal fator 1 e o eixo y é o principal fator 2, podemos observar que o disco 1 realmente se aproxima do eixo do principal fator 1 e a CPU idem ao eixo y. O disco 2, como analisado no PCA representa pouco os dados, e o gráfico 10 evidencia esse fato, o ponto disco 2 está disperso do eixo x e do eixo y, confirmando os resultados obtido pelo PCA.

4 Dados -Questão A

4.1 Software Utilizados

Para o desenvolvimento do trabalho utilizamos o software editor de planilhas OpenOffice Calc e o Calcular autovalores e autovetores de uma matriz do Wolfram Alpha [2]. Todos os dados estão disponíveis no repositório *https://github.com/mateusstp/avaliacaodesempenhosistemas*.

4.2 Dados do problema e dados estatísticos

Tabela 2 – Dados do problema e dados estatísticos

	Coleta Real				Normalizado			Saídas		
Número	**CPU**	**Disk 1**	**Disk 2**	**TR ID**	**CPU**	**Disk 1**	**Disk 2**	**y1**	**y2**	**y3**
1	116,824	9,000	9,000	18	-0,732	-1,570	-1,357	0,544	1,942	-0,879
2	64,383	7,000	9,000	37	-1,048	-1,644	-1,357	0,696	2,162	-0,695
3	35,403	7,000	9,000	58	-1,223	-1,644	-1,357	0,801	2,248	-0,585
4	104,409	8,000	12,000	77	-0,807	-1,607	-1,243	0,498	1,990	-0,755
5	119,793	9,000	8,000	19	-0,714	-1,570	-1,395	0,557	1,940	-0,919
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
198	139,368	60,000	40,000	57	-0,596	0,320	-0,183	0,628	0,057	0,306
199	136,998	68,000	37,000	201	-0,610	0,616	-0,297	0,852	-0,167	0,292
200	149,211	69,000	38,000	17	-0,537	0,653	-0,259	0,802	-0,242	0,282
$\sum x$	47640,823	10275,000	8969,000	-	0,000	0,000	0,000	0,000	0,000	0,000
$\sum x^2$	16822813,796	672893,000	541135,000	-	199,000	199,000	199,000	207,390	348,108	41,503
Média	238,204	51,375	44,845	-	0,000	0,000	0,000	0,000	0,000	0,000
Variância	27510,420	728,718	698,091	-	1,000	1,000	1,000	1,042	1,749	0,209
Desvio Padrão	165,863	26,995	26,421	-	1,000	1,000	1,000	1,021	1,323	0,457

Tabela 3 – Dados estatísticos

	CPU	**Disk 1**	**Disk 2**
Coeficiente de Variação	0,696	0,525	0,589
Valor Máximo	507,450	85,000	92,000
Valor Mínimo	23,597	5,000	7,000
Range	483,853	80,000	85,000
1º Quartil	104,439	33,000	26,250
2º Quartil	151,625	63,000	39,000
3º Quartil	418,052	72,000	68,000
Mediana	151,625	63,000	39,000

5 Gráficos - Questão B

5.1 **Histogramas**

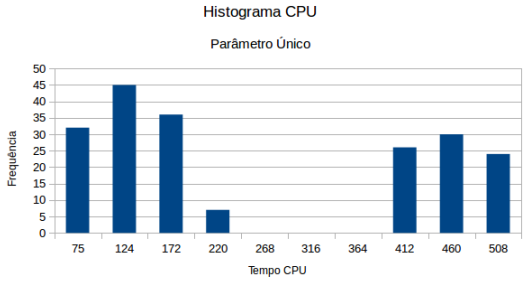

Figura 1 – Histograma CPU com 10 Classes

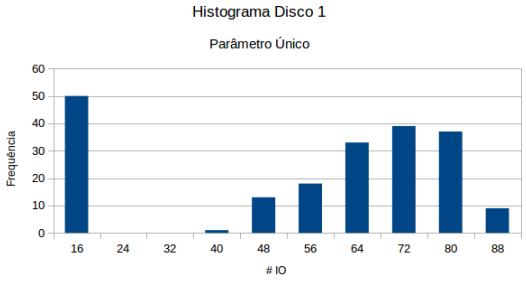

Figura 2 – Histograma Disco 1 com 10 Classes

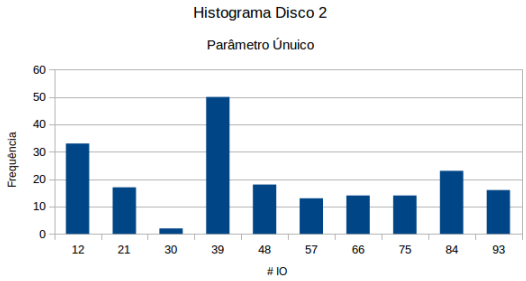

Figura 3 – Histograma Disco 2 com 10 Classes

5.2 CDF

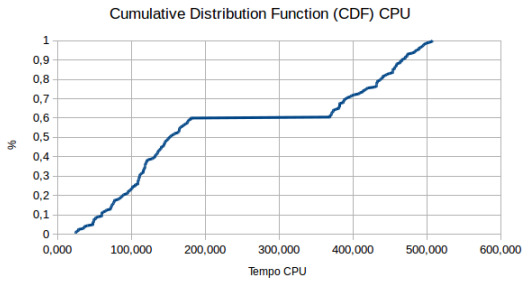

Figura 4 – CDF CPU

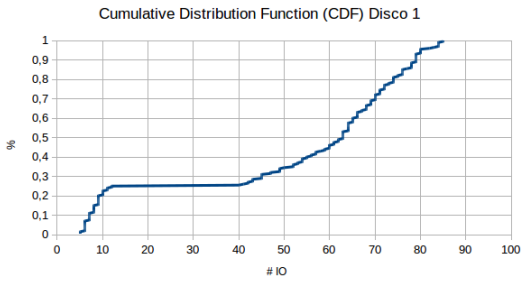

Figura 5 – CDF Disco 1

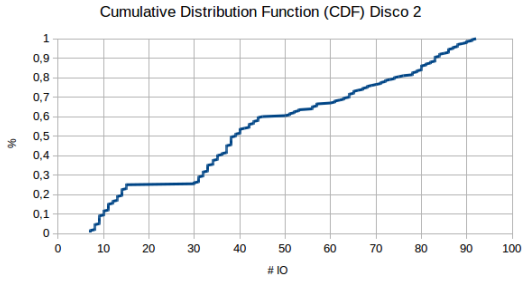

Figura 6 – CDF Disco 2

5.3 Box Plot

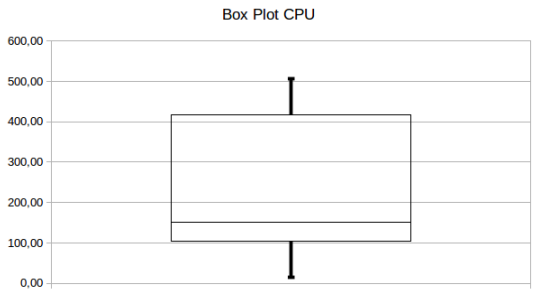

Figura 7 – Box plot CPU

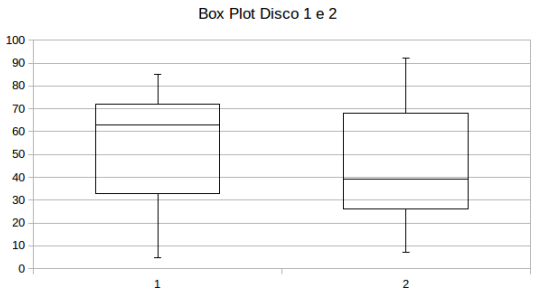

Figura 8 – Box plot Disco 1 e 2

6 Análise do Componente Principal - Questão C

Tabela 4 – Autovetores

v1	**v2**	**v3**
-0,596	0,485	-0,640
-0,496	-0,849	-0,181
-0,631	0,209	0,747

Tabela 5 – Matriz de Correlação

1,00	0,465	0,916
0,465	1,00	0,626
0,916	0,626	1,00

Tabela 6 – Correlação entre variáveis

Correlação	$(CPU, Disco1)$	$(CPU, Disco2)$	$(Disco1, Disco2)$
$R_{(i,j)}$	0,465	0,916	0,626

Tabela 7 – Porcentegem PCA

	CPU	**Disk 1**	**Disk 2**
Porcentagem de cada fator	34,74%	58,31%	6,95%

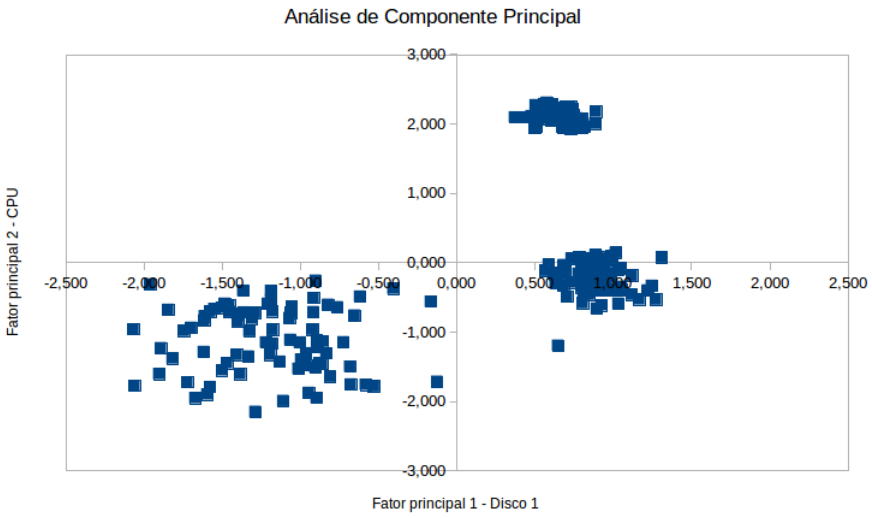

Figura 9 – PCA

7 Correlação PCAs e Componentes - Questão D

Tabela 8 – Correlação PCA e Componentes

	y1	**y2**
$R_{(pcu,yi)}$	-0,799	-0,938
$R_{(d1,yi)}$	-0,902	-0,189
$R_{(d2,yi)}$	-0,882	-0,865

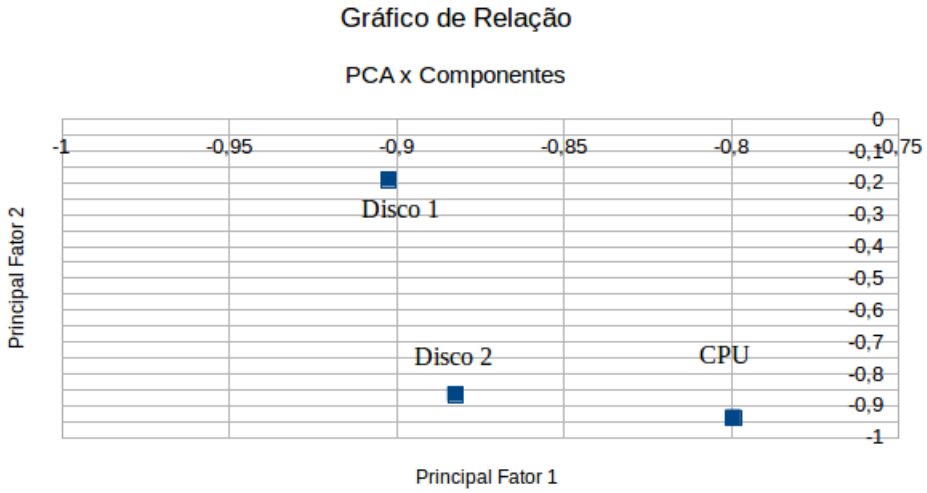

Figura 10 – Correlação PCA e Componentes

Referências

[1] **JAIN**, Raj. The Art of Computer System Performance Analysis. *EUA: John Wiley & Sons, 1991.*

[2] **Equipe IGM**. Cálculo de autovalores e autovetores. *disponível em http://www.igm.mat.br/aplicativos/index.php?option=com_content&view=article&id=754:auto-valores-vetores&catid=101:edo-segunda-ordem.*