



PRINCIPAL COMPONENT ANALYSIS - PCA

Curso: Engenharia de Computação

2º semestre de 2015

Disciplina: Avaliação de Desempenho de Sistemas

Data: 06/02/16

Professor: Alexandre Magno de Sousa

VALOR: 5 PONTOS EXTRA

Aluno: _____

Data de entrega: 12/02/16

- 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:
 - 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.
 - apresente gráficos para cada *feature* tais como: histograma de único parâmetro, *Cumulative Distribution Function* (CDF) e Box Plot.
 - realize a Análise do Componente Principal (PCA) seguindo todos os passos apresentados em sala de aula, não se esqueça de construir os gráficos.
 - 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

$$R(i, j) = \frac{\frac{1}{n-1} \sum_k (i_k - \bar{i})(j_k - \bar{j})}{s_i \times s_j}$$

onde n representa o número total de componentes registrados, \bar{i} e \bar{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.

- mostre os resultados, faça observações e apresente conclusões.

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

Tabela 1: Tabela de análise PCA.

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.