Universidade Federal de Campina Grande
Departamento de Sistemas e Computação
Curso de Bacharelado em Ciência da Computação
Disciplina: Análise de Dados 1

Problema 02 - Sumarizando seus dados

Para este problema, usaremos dados reais de ociosidade e uso de máquinas de alguns laboratórios do DSC, sendo dois laboratórios de pesquisa (GMF e LSD) e dois laboratórios de alunos de graduação (LCC e LCC2). A máquina é considerada ociosa quando não há atividade do usuário interagindo na mesma, ou seja, o teclado e o *mouse* não são utilizados por um certo período de tempo.
Os dados possuem as seguintes colunas:

- **intervalo** *[integer]*: tempo em segundos em que a máquina passou em um determinado estado (ociosa ou ocupada).
- **ociosa** *[logical]*: indica o estado da máquina correspondente ao intervalo de tempo indicado, sendo *TRUE* quando a máquina está ociosa ou *FALSE* quando está em uso.
- **maquina** *[character]*: identificador da máquina.
- **laboratorio** *[character]*: identificador do laboratório ao qual a máquina pertence.

Os dados podem ser acessados aqui (está zipado, vc tem que deszipar antes de usar no R).

Neste problema, teremos dois tipos de resposta. Uma em formato de *script R* e outra em formato texto *pdf*. Os arquivos de resposta devem ter os seguintes nomes:

- *seunome-prob02.R* para respostas no formato de scripts *R*
- *seunome-prob02.pdf* para respostas discussivas em formato texto.

Problema:

1) Sumarize os dados coletados e use essa sumarização para extrair informações interessantes sobre o uso das máquinas nos 4 laboratórios monitorados.

Dicas:

1. Entenda os dados e suas escalas
2. Entenda o formato das distribuições de cada variável
3. Calcule índices de tendência central e dispersão que julgar mais adequados
4. Use gráficos como boxplot, histogramas e scatterplots para melhor visualizar os dados

e entender formatos de distribuições, tendências, outliers…

5. Lembre que você pode agrupar os dados de diversas formas e que estatísticas e gráficos podem ser calculadas e gerados para grupos diferentes de dados. Exemplos: agrupar por laboratório, por máquina. Os itens 2, 3 e 4 podem ser feitos para cada agrupamento considerado. *Falando em R, tem uma biblioteca bem bacana (mais sofisticada) chamada plyr. Talvez o ddply seja uma alternativa bem sofisticada ao aggregate e ao tapply.*
6. Será que tem algo que possa ser calculado considerando as variáveis nominais/categóricas?
7. Pense bem numa boa organização para o documento de sumarização de dados, não mostre simplesmente numeros e gráficos de qualquer jeito - o bom analista se preocupa com o fluxo de ideias em um relatório, além de ser recomendado o uso de gráficos, tabelas/números.
8. Será que existem outliers em seus dados? Se você achar que existem e que devem ser removidos, remova-os, mas tenha um argumento de justificativa para remoção de outliers em seu relatório.
9. Seja curioso… Quais são as máquinas mais ocupadas? E as mais ociosas? Tem outras perguntas que você pode se fazer…

Obs: *consultar os comandos ?png e ?dev.off para salvar as imagens.*

2) Suponha que você deve fazer uma análise de qual laboratório as pessoas mais trabalham / estudam (ou seja, as máquinas ficam mais tempo ocupadas). Qual laboratório você indicaria? Discuta como você chegou à sua conclusão.

Desafio (opcional):

1 Descrição

Esse desafio consiste em duas partes: (i) implementação em R da solução de um problema que usa medidas estatísticas aprendidas em aula; (ii) discussão das decisões estatísticas definidas no algoritmo.

1.1 Implementação

Você está procurando um computador do DSC para rodar um serviço de monitoramento de mensagens no Twitter. Você avalia o computador sob dois critérios: disponibilidade e volatilidade. A disponibilidade é o total de tempo em que o computador permanece com o atributo “ociosa” com o valor “TRUE”, quanto mais disponível a máquina for mais tráfego de

mensagens ela poderá monitorar. Por outro lado, a volatilidade é a quantidade de vezes em que o computador mudou o atributo “ociosa”, quanto mais volátil a máquina for mais vezes você precisará acessá-la para verificar manualmente o estado do monitoramento e reiniciar o serviço.

Primeiramente, aplique a *Rule of Thumb* para eliminar computadores com disponibilidade e volatilidade destoantes, i.e, possíveis outliers. Após isso, em uma fase de pre-processamento realizada em cada laboratório, elimine os computadores que apresentem ambos disponibilidade menor que o 55 percentil e volatilidade maior que o 55 percentil. Do conjunto de todos os computadores que satisfazem essa restrição, você está interessado em identificar aquele que apresenta maior disponibilidade e calcular quanto porcento ele é mais disponível e quanto porcento ele é menos volátil que a média e a mediana dos dados filtrados.

O desafio será avaliado pela corretude da resposta gerada segundo a descrição acima.

1.2 Discussão das decisões

Além do script, você deve discutir sobre: (i) o que você mudaria no algoritmo para que ele fizesse uma melhor escolha e (ii) se você acha que esse algoritmo pode ser generalizado para outras bases de dados.

2 Instruções para entrega

Você deve entregar um script *R* que recebe como entrada um arquivo CSV de atividade das máquinas (no formato dos dados de exemplo do lab). Exemplo de execução do script: $ Rscript meuscript.R dados.csv. A implementação deve seguir a descrição do problema e gerar como saída em um arquivo em formato texto com o nome da máquina escolhida, o laboratório em que ela está e o ganho que ela proporciona em termos de volatilidade e disponibilidade em relação à média e a mediana. Além disso, deve ser entregue um arquivo *pdf* com a discussão das decisões.