Trabalho Prático

Análise de Dados em Informática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2021/2022

---

1. Objetivos

Objetivo Geral:

- Análise Exploratórias de Dados
- Análise Inferencial
- Correlação e Regressão

Objetivos específicos:

- Definir a metodologia de trabalho
- Análise e discussão dos resultados com recurso ao R
- Escrita de Relatório Técnico com a Análise de Dados

2. Calendarização

Lançamento das propostas de trabalhos: até 20 de março de 2021
Entrega do trabalho: até **23 de abril de 2022** (23:55)
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O grupo (máx 3 elementos) deve ser o mesmo nas 2 iterações do Trabalho Prático.
- Deverá ser usado o R como ferramenta de suporte ao tratamento de dados.
- A **data final de ENTREGA** do 1º Trabalho Prático é **23 de abril de 2022,** no moodle. Independentemente deste prazo, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.
- A entrega do trabalho consta de um Relatório Técnico. Deverá submeter todos os documentos num ficheiro compactado. O zip file deve conter:
    - Relatório Técnico em pdf
    - dados utilizados em formato csv
    - script completo (e comentado) do código criado em R para resolver o problema
- O nome do ficheiro deverá seguir a seguinte notação:

  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_AIM_3DA_7777777_8888888_9999999.zip.**
- Trabalhos cuja designação não respeite a notação indicada, **serão penalizados em 10%.**
- **A entrega do trabalho deverá ser submetida no moodle até à data de entrega definida. Não serão aceites trabalhos fora do prazo.**
- A defesa e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação. A defesa e discussão serão realizadas em grupo com questões direcionadas a cada elemento individualmente.
- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas teórico-práticas.

3.1. Relatório Técnico

No Relatório Técnico deverão ser documentadas todas as fases da metodologia de trabalho seguida, contextualização do tema, exploração, preparação dos dados, análise e discussão dos resultados e conclusões.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos e as ponderações especificadas na tabela:

1. Contextualização e enquadramento teórico, motivação e objetivos (Introdução)
2. A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados e as conclusões alcançadas
3. Organização, qualidade da escrita, apresentação e clareza do relatório
4. A defesa e discussão
5. Participação individual de cada um dos elementos

Tabela 1 – Grelha de avaliação da 1ª iteração do Trabalho Prático

Introdução	**10%**
Análise do funcionamento dos servidores VPN do DEI	
Exercício 1	15%
Exercício 2	15%
Análise de Desempenho de Métodos Heurísticos na resolução do problema de Escalonamento	
Exercício 1	25%
Exercício 2	15%
Conclusões	**10%**
Estrutura e organização do artigo	**10%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho (que poderá ser por videoconferência), a participação de cada um dos elementos do grupo na concretização dos objetivos do trabalho e do grupo. **Os elementos ausentes não terão classificação.**

4. Descrição do Trabalho

Na realização da 1ª iteração do Trabalho Prático pretende-se que os alunos desenvolvam o processo de Análise Exploratória de Dados [1], Análise Inferencial, Correlação e Regressão em duas áreas de intervenção: a Análise de Fiabilidade e Análise de Desempenho de Algoritmos. São propostos dois problemas a analisar: Análise do funcionamento dos servidores VPN do DEI e a Análise de Desempenho de Métodos Heurísticos na resolução do problema de Escalonamento.

4.1. Análise do funcionamento dos servidores VPN do DEI

O DEI mantém vários servidores de rede privada virtual (VPN - Virtual Private Network). Estes serviços permitem aos utilizadores localizados em redes exteriores ao DEI criar uma ligação de rede virtual no seu posto de trabalho equivalente a uma ligação física às redes internas do DEI. Com esta ligação, os serviços das redes do DEI que não estão acessíveis publicamente passam a estar disponíveis.
Como acontece com todos os serviços de rede do DEI, é mantido um registo das sessões VPN estabelecidas pelos utilizadores. Com base neste registo de atividades pretendem-se vários tipos de análise ao funcionamento destes serviços e a obtenção de conclusões relevantes para a respetiva administração.
O registo de atividades a ser objeto de análise é fornecido sob a forma de um ficheiro de texto que por motivos de privacidade foi anonimizado (os nomes dos utilizadores e os endereços de rede foram removidos). Cada sessão é apresentada sob a forma de uma linha de texto, em sequência temporal, contendo os seguintes elementos: servidor, protocolo de VPN, data de início da sessão, hora de início da sessão, hora do fim da sessão, e respetiva duração em minutos.

Extrato exemplificativo:

```
vsrv17 PPTP 2016-12-25 11:28 18:53 (445 min.)
vsrv8 SSTP 2016-12-25 18:24 19:39 (75 min.)
vsrv17 SSTP 2016-12-25 19:45 19:45 (0 min.)
vsrv17 PPTP 2016-12-25 19:05 19:57 (52 min.)
vsrv17 SSTP 2016-12-25 16:30 20:43 (253 min.)
vsrv17 PPTP 2016-12-25 19:56 20:56 (60 min.)
vsrv16 SSTP 2016-12-25 19:39 21:01 (82 min.)
vsrv8 PPTP 2016-12-25 20:57 21:16 (19 min.)
vsrv16 PPTP 2016-12-25 21:25 21:48 (23 min.)
vsrv11 SOFTETHER 2016-12-25 21:10 22:21 (71 min.)
```

O ficheiro fornecido (**`vpnsessionsfile.txt`**) tem um total de 66383 linhas, abrangendo um período de utilização superior a um ano, de **2016-12-25** até **2018-02-22** . Existem cinco servidores: **vsrv8**; **vsrv10**; **vsrv11**; **vsrv16**; **vsrv17**. Todos os servidores são multiprotocolo, isto é, suportam vários tipos de protocolo de **VPN**: **PPTP**; **SSTP**; **SOFTETHER**; **OPENVPN L2**; **OPENVPN L3**.

Definições:

a) **Falha** - embora a razão de uma sessão ter uma curta duração possa ter diversas origens, para o efeito deste estudo considera-se que qualquer sessão com duração igual ou inferior a um minuto representa uma falha.

b) **Número de falhas simultâneas** - Em cada minuto t_0, define-se o número de falhas simultâneas como sendo o número de falhas que inicializaram ou que finalizaram no minuto t_0.

c) **Acesso** - Um acesso é uma sessão com duração superior a um minuto.

d) **Número de acessos simultâneos** - Em cada minuto t_0, define-se o número de acessos simultâneos como sendo o número de acessos que inicializaram antes do minuto t_0 e terminaram depois do minuto t_0.

No relatório deve responder, justificar e comentar as suas respostas:

1. Responda às seguintes questões na forma que ache mais apropriada (Use o conjunto completo de dados):
   a) Quantos acessos teve cada servidor?
   b) Quantas falhas teve cada servidor?
   c) Quantas vezes o servidor "**x**" usa o protocolo "**y**"?
   d) Determine as médias, medianas e desvios padrão mensais de acessos de cada servidor (apenas meses completos)

2. Com os dados relativos ao mês de dezembro de 2017 (todos os servidores e todos os protocolos):
   a) Efetue um gráfico que nas abcissas represente o tempo, e nas ordenadas o número de falhas simultâneas e o número de acessos simultâneos.
   b) Efetue um diagrama de caixa de bigodes do número diário de falhas simultâneas, para cada servidor (o número diário de falhas simultâneas é o número total de falhas simultâneas que ocorreram nesse dia). Indique quantos *outliers* existem por servidor.
   c) Verifique se há correlação entre o número de falhas simultâneas e o número de acessos simultâneos, no dia 11 de dezembro de 2017 das 12:00 às 14:00.

4.2. Análise de Desempenho de Métodos Heurísticos na resolução do problema de Escalonamento

Um problema de escalonamento é geralmente caracterizado por um conjunto $T = \{T_1, T_2, \ldots, T_n\}$ de tarefas a executar, num conjunto de máquinas $M = \{M_1, M_2, \ldots, M_m\}$. Nesta classe de problemas pretende-se encontrar uma solução que corresponda à melhor sequência de processamento das tarefas do conjunto T, numa ou mais máquinas do conjunto M, de modo que todas as tarefas sejam executadas, respeitando as limitações das máquinas e eventuais restrições adicionais impostas ao problema.
O problema de escalonamento [2] é conhecido por ser um problema de complexidade NP-hard ou seja, a partir de uma certa dimensão torna-se impraticável encontrar a solução ótima de forma eficiente (ou seja, resolver o problema de forma exata). Uma das métricas de desempenho (funções objetivo) mais usada para se calcular o valor da solução é o *makespan*, C_{max}. Onde,

C_{max} = max{C_i}, em que C_i é o instante de tempo da conclusão da tarefa T_i.

Note-se que, a resolução exata fornece o plano com o *makespan* mínimo (quanto menor for o *makespan* mais eficaz é a solução). Geralmente, recorre-se para resolver instâncias de grande dimensão a técnicas inspiradas na Inteligência Artificial – Meta-Heurísticas (MH), que resolvem um problema de escalonamento de forma satisfatória, mas que, geralmente, não garantem a solução ótima.

1. O ficheiro (**`mspandata.csv`**) contém resultados relativos a 80 instâncias do problema de escalonamento. As instâncias encontram-se identificadas com números de 1 a 80, tendo sido apresentadas por ordem crescente de tamanho (por exemplo: o problema 1 é o problema com menor dimensão e o problema 80 é o problema de maior dimensão). Para cada problema conhece-se o valor do *makespan* ótimo identificado por OPT e o *makespan* obtido por cada MH identificadas por MH1, MH2 e MH3:

   a) Efetue um gráfico tal que: o eixo das abcissas contém os problemas e o eixo das ordenadas contém o valor do *makespan* das três MH's para cada instância (use cores para distinguir as MH).

   b) Considere apenas os dados relativos às 10 instâncias com menor tamanho. Haverá diferenças significativas entre o desempenho das três técnicas?

   c) No caso de a resposta da alínea anterior ser positiva, identifique qual a técnica mais eficaz.

   d) Responda às duas perguntas anteriores, mas usando os dados das 20 instâncias com maior dimensão.

   e) Por vezes usa-se o *makespan* normalizado, C_{norm} para avaliar o desempenho de uma MH. Com,

   $$C_{norm} = (C_{MH} - C_{OPT})/C_{OPT}$$

   onde C_{MH} é o *makespan* de uma MH e C_{OPT} é o *makespan* ótimo. Repita todas as alíneas anteriores usando o *makespan* normalizado (em vez do *makespan*).

   f) Comente os resultados obtidos usando as duas medidas de desempenho usadas (*makespan* e o *makespan* normalizado).

   g) Use os dados relativos às 70 instâncias de menor dimensão. Supondo que o tamanho do problema é a variável preditora e que o *makespan* normalizado é a variável resposta determine, para cada MH, a reta de regressão linear.

   h) Verifique se os pressupostos sobre os resíduos são verificados (normalidade, homocedasticidade e independência).

   i) Comente os resultados obtidos nas duas alíneas anteriores.

2. Para avaliar a eficiência das MH é usual considerarmos o tempo de processamento na obtenção da solução (CPU time). Quanto menor o tempo de execução de um método mais eficiente será a técnica de otimização. O ficheiro (***`Cpu.run.time.csv`***) contém o tempo de processamento (em segundos) das três MH (MH1, MH2 e MH3) na resolução das 50 instâncias com menor dimensão.

   a) Construa um boxplot que contenha os tempos de processamento de cada MH na resolução de cada instância.

   b) Verifique se existem diferenças significativas nos tempos médios de processamento entre as três técnicas.

   c) No caso da resposta da alínea anterior ser positiva, identifique qual a MH mais eficiente.

   d) Determine a matriz de correlação entre os tempos de processamento de cada MH e interprete os resultados.

3. Efetue um resumo dos resultados e das conclusões, obtidos neste trabalho, que considera mais importantes

5. Referências Bibliográficas

[1]. HEUMANN, C., M. SCHOMAKER and SHALABH, Introduction to statistics and data analysis, Springer International Publishing, 2016.

[2]. MADUREIRA, ANA MARIA, Aplicação de Meta-Heurísticas ao Problema de Escalonamento em Ambiente Dinâmico de Produção Discreta, PhD Thesis at Universidade do Minho, 2003.