Metodologias Experimentais em Informática - Relatório de Experiência 2

Edgar Antunes, Guilherme Tavares, Henrique Silva

dezembro 2021

Resumo

Neste trabalho pretende-se analisar as diferenças de desempenho para dois algoritmos, cujo objetivo é determinar o número mínimo de blocos horários em que é possível alocar um determinado número de exames. Este estudo engloba uma análise exploratória de dados, recorrendo a técnicas gráficas, e uma análise de regressão dos dados experimentais na expectativa de identificar tendências.

Palavras Chave: Testes de Hipóteses - Intervalos de Confiança - Algoritmos de *backtracking*

1 Configuração das Experiências

Para minimizar as variáveis de ambiente todas as experiências foram executadas no mesmo computador.

A escolha dos valores para o número de exames foi tomada para haver alguma abrangência, para tal os valores estão compreendidos entre $n = 10$ e $n = 50$ com um incremento uniforme, sendo o limite de 50 exames por razões de performance.

Relativamente à escolha das probabilidades, foram escolhidos valores entre $p = 0.05$ e $p = 0.33$, sendo que não foram considerados valores superiores por serem computacionalmente intratáveis.

Foi definido também um limite máximo de execução de 20 minutos por experiência, pois tendo em conta os valores que fomos obtendo revelou-se ser um limite adequado que nos permitiu obter resultados em todas as experiências menos uma (a mais complexa, com valores de $n = 50$ e $p = 0.33$).

Para cada par de n e p, corremos 20 vezes a experiência, guardando todos os valores obtidos para serem utilizados.

Todos os *scripts* necessários para a reprodução das experiências estão disponíveis na pasta referente ao código.

2 Testes de Hipóteses

2.1 Diferença de Performance entre os dois algoritmos

Hipótese Nula 1 (H0): *Não há diferença significativa de performance entre os algoritmos.*

Hipótese 1 (H1): *Há uma diferença significativa de performance entre os algoritmos.*

Para testar esta hipótese, iremos comparar os valores de *runtime* obtidos nas experiências efetuadas fixando $n = 50$. Em primeiro lugar, efetuamos o teste de Shapiro para verificar a normalidade dos dados, presunção necessária para a aplicação de testes paramétricos.

```
    Shapiro-Wilk normality test

data:  code1_size_50$runtime
W = 0.71357, p-value = 1.13e-12
```

Como o valor de p obtido no teste de shapiro $p = 1.13e - 12 < 0.05$, então rejeitamos a hipótese nula, pelo que afirmamos que os dados não estão normalizados. Deste modo, foi feito o teste de Wilcoxon, a escolha deste teste não paramétrico foi baseada no uso de amostras emparelhadas.

```
   Wilcoxon rank sum test with continuity correction

data:  code1_size_50$runtime and code2_size_50$runtime
    W = 4861, p-value = 0.7325
alternative hypothesis: true location shift is not equal to 0
```

Após o teste de Wilcoxon, obtemos um valor de $p = 0.7325 > 0.05$ pelo que rejeitamos a hipótese nula, ou seja, aceitamos a hipótese alternativa de que há uma diferença de performance significativa entre os algoritmos.

2.2 Escalabilidade dos algoritmos

Hipótese Nula 2 (H0): *O algoritmo 1 não escala melhor que o algoritmo 2.*

Hipótese 2 (H1): *O algoritmo 1 escala melhor que o algoritmo 2.*

Para esta hipótese, recolhemos os coeficientes da aplicação dos modelos de regressão efetuados durante a fase de análise exploratória deste trabalho. As experiências utilizadas foram aquelas em que fixávamos os valores de p, e recolhemos o coeficiente associado a cada experiência, criando assim um *dataset* de coeficientes associado a cada algoritmo.

Coeficientes Algoritmo 1	Coeficientes Algoritmo 2
0.1657128	0.1959227
0.3440161	0.3747055
0.3858081	0.4247491
0.4468239	0.4671112
0.4256178	0.4454711

De seguida, efetuamos o teste de Shapiro para verificar a normalidade dos dados, presunção necessária para a aplicação de testes paramétricos.

```
    Shapiro-Wilk normality test

data:  coefficcients code1
W = 0.84789, p-value = 0.1879

    Shapiro-Wilk normality test

data:  coefficcients code2
W = 0.81495, p-value = 0.1067
```

Como o valor de p obtido no teste de Shapiro para os coeficientes do código1 é $p = 0.1879 > 0.05$ e os coeficientes do código2 $p = 0.1067 > 0.05$, então não rejeitamos a hipótese nula, ou seja, os dados seguem uma distribuição normal.

Assim, uma vez que se tratam de amostras desemparelhadas, efetuámos o T-teste de Welch.

```
    Welch Two Sample t-test

data:  coefficcients1 and coefficcients2
t = -0.39977, df = 7.9947, p-value = 0.3499
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf 0.10224
sample estimates:
mean of x mean of y
0.3535958 0.3815919
```

Experiências com seeds aleatórias	Experiências com seeds selecionadas
9.164137e-06	1.581696e-05
7.242797e-02	1.156728e-03
3.730508e-01	7.978762e-01
1.932655e+02	5.217649e+01
5.451453e+02	4.704046e+02
1.305576e-05	1.022065e-05
2.254919e-02	1.178920e-03
2.247131	7.811189e-01
3.099299e+02	5.280743e+01
5.385406e+02	4.704768e+02

Tabela 1: Desvios padrão das experiências, para cada combinação de número de exames e probabilidade de exames com alunos em comum

Como se pode verificar pelos resultados do teste, o valor de p obtido permite-nos refutar a hipótese nula; deste modo, podemos afirmar que o algoritmo 1 escala melhor que o algoritmo 2, isto é, tem um crescimento menor em relação ao tamanho dos *datasets*, para os mesmos valores de n e p.

2.3 Impacto das *seeds* na performance dos algoritmos

Hipótese Nula 3 (H0): *As seeds das experiências escolhidas não afetam a dispersão dos dados.*

Hipótese 3 (H1): *As seeds das experiências escolhidas afetam a dispersão dos dados.*

Para esta hipótese, recolhemos os desvios padrão dos tempos de execução em cada uma das experiências realizadas na primeira versão da primeira meta, onde foram utilizados, quer os inputs, quer as experiências, os mesmos valores das *seeds*, bem como os tempos de execução em cada uma das experiencias realizadas na última versão da primeira meta, onde para cada experiência realizada e para cada input gerado, as *seeds* tinham valores aleatórios.

Para decidir que teste aplicar para a rejeição, ou não, da hipótese em teste, efetuámos, em cada um dos dados, o teste de Shapiro

```
    Shapiro-Wilk normality test

data:  sd_RandomSeeds
W = 0.71926, p-value = 0.001509
```

```
    Shapiro-Wilk normality test

data:  sd_selectedSeeds
W = 0.5783, p-value = 3.052e-05
```

Como $p - value < 0.05$, neste teste de Shapiro, concluímos que os dados não seguem uma distribuição normal.

Assim, uma vez que se tratam de amostras desemparelhadas e que não seguem uma distribuição normal, efetuámos o teste de Mann-Whitney.

```
    Wilcoxon rank sum exact test

data:  sd_RandomSeeds and sd_SelectedSeeds
W = 55, p-value = 0.7394
alternative hypothesis: true location shift is not equal to 0
```

Sendo que $p - value = 0.7394 > 0.05$, podemos rejeitar H0 para um nível de confiança de 0.05.

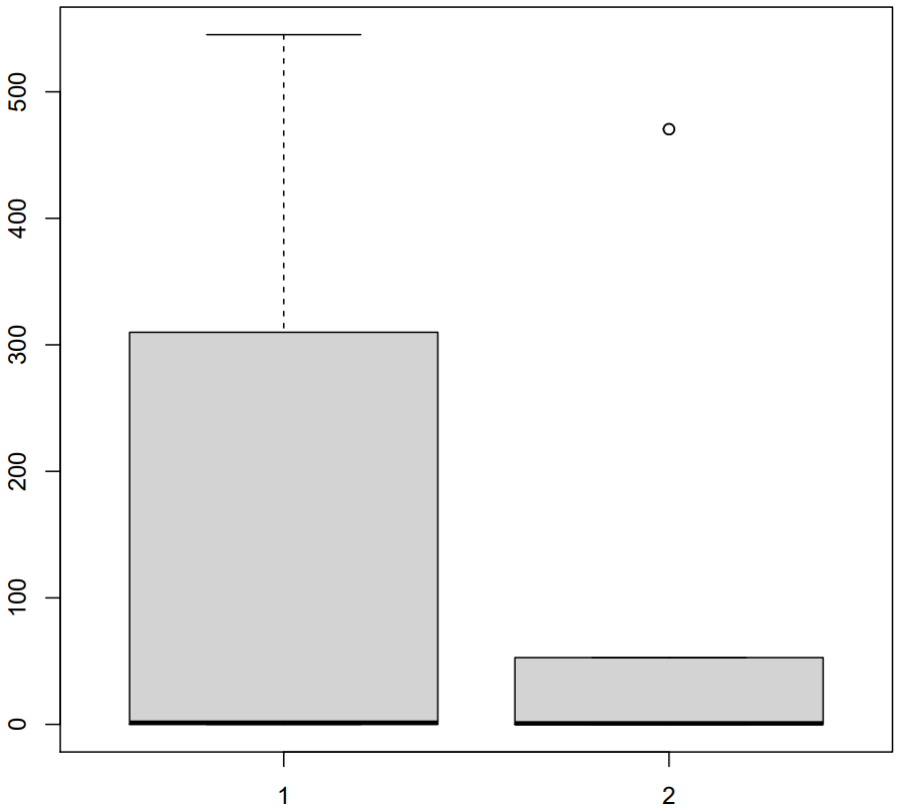

Figura 1: Valor dos desvios padrão para as experiências com seeds aleatórias (1) e seeds selecionadas (2)