Trabalho Prático

Análise de Dados em Informática

Análise de Desempenho

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2019/2020

---

1. Objetivos

Objetivo Geral:

- Análise de Desempenho de técnicas de aprendizagem automática

Objetivos Específicos:

- Definir a metodologia de trabalho
- Análise e discussão dos Resultados com recurso ao R
- Escrita de artigo científico

2. Calendarização

Entrega do trabalho: até **13 de junho de 2020 pelas 23:55**
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O grupo deve ser o mesmo em todos os trabalhos práticos.

- Deverá ser usada a ferramenta R.

- A **data final de ENTREGA** do trabalho é **13 de junho de 2020 pelas 23:55**, no moodle.
  Independentemente destes prazos, 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 artigo científico (**máx. 8 páginas**) conforme *template* disponibilizado no moodle, apresentação *powerpoint* com resumo do trabalho realizado, entre outros. Deverá submeter todos os documentos num ficheiro compactado. O zip file deve conter:
  - artigo científico em pdf
  - dados utilizados em formato csv
  - script completo (e comentado) do código criado em R para resolver o problema
  - apresentação PowerPoint com resumo do artigo para 10 minutos (ppt)

- O nome do ficheiro zip 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_AMD_3AD_7777777_8888888_9999999.zip.**

- Trabalhos cujo nome 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 apresentação, **em formato de comunicação (10 minutos),** 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 avaliação do trabalho será realizada pelo docente das aulas TP.

- 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. *Artigo Científico*

No Artigo Científico (máx. 8 páginas) deverão ser documentadas todas as fases da metodologia de trabalho seguida, contextualização do tema, exploração, preparação dos dados, análise e discussão dos resultados, conclusões e referências bibliográficas.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos:

- Revisão do estado da arte (algoritmos de aprendizagem automática e análise de desempenho);
- Desenvolvimento de modelos de *Machine Learning*;
- A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados e as conclusões alcançadas;
- Organização, qualidade da escrita, apresentação e clareza do artigo científico;
- A apresentação numa aula e discussão;
- Participação individual de cada um dos elementos.

Contextualização (Abstract, Introdução, estado da arte)	**10%**
Análise de desempenho de técnicas de aprendizagem	**70%**
Conclusões	**10%**
Apresentação e Discussão	**10%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. No momento da defesa do trabalho será validada a participação de cada um dos elementos do grupo na concretização dos objetivos do trabalho e do grupo.

4. Descrição do Trabalho

O objetivo principal deste trabalho consiste na aplicação de algoritmos de aprendizagem automática na exploração de dados e respetiva comparação dos mesmos usando os testes estatísticos mais adequados. Deve ser produzido um artigo científico (português ou inglês), conforme *template* indicado, com o estado da arte sobre os diferentes algoritmos, os modelos desenvolvidos, os resultados obtidos, a análise e discussão dos resultados e as conclusões.

Pretende-se que façam a análise salarial da população masculina de uma dada região, com base nos atributos abaixo descritos, através de modelos de classificação/regressão usando os algoritmos estudados: regressão linear, árvores de decisão, k-vizinhos-mais-próximos e redes neuronais.

O conjunto de dados a analisar neste trabalho diz respeito a salários e outras informações para 3000 trabalhadores do sexo masculino. Pretende-se que explorem as relações entre salário e os restantes atributos deste conjunto de dados:

Atributo	Descrição
Ano	Data de registo da informação
Idade	Idade do funcionário
Sexo	Sexo do funcionário
Estado	Estado civil
Raça	Etnia
Grau	Nível de estudos do funcionário
Emprego	Tipo de emprego
Saude	Nível de saúde do funcionário
Seguro	Funcionário com seguro de saúde (sim/não)
Salario	Salário dos funcionários

1. Comece por carregar o ficheiro ("**dados_emprego.csv**") para o ambiente do R, verifique a sua dimensão e obtenha um sumário dos dados.

2. Faça um estudo da regressão linear entre a variável dependente (**Salario**) e a variável independente (**Idade**):
   a. Calcule a correlação entre as variáveis **Salário** e **Idade**;
   b. Encontre a reta de regressão linear entre a variável dependente (**Salário**) e a variável independente (**Idade**);
   c. Verifique as condições sobre os resíduos (normalidade, independência e homocedasticidade).

3. Encontre o modelo de regressão linear generalizado onde a variável dependente é o "**Salario**" e as variáveis independentes são a "**Idade**", o "**Grau**" e o "**Emprego**". Com base no modelo encontrado, estime:
   a. Uma pessoa com 32 anos, com o emprego de "Industrial", para todos os diferentes graus de escolaridade;
   b. Uma pessoa com 32 anos com o "12º ano escolaridade", para todos os diferentes tipos de emprego.

4. Derive um novo atributo **Nivel**, discretizando o atributo **Salario** em duas classes: Alto e Baixo, usando como valor de corte a mediana.

5. Faça uma Análise Exploratória de Dados, usando os gráficos apropriados, de modo a analisar os vários atributos (numéricos e categóricos) do conjunto de dados.

6. Usando o método ***k-fold cross validation*** estude a capacidade preditiva de alguns métodos de previsão relativamente ao novo atributo **Nivel**:
   a. Um modelo árvore de decisão;
   b. Uma rede neuronal. Deve avaliar o desempenho para diferentes configurações;
   c. Compare as soluções obtidas para as medidas de avaliação ***accuracy***, **precision**, ***recall*** e ***F1;***
   d. Verifique se existe diferença significativa no desempenho dos diversos algoritmos (use um nível de significância de 5%). Identifique a técnica de aprendizagem automática que apresenta melhor desempenho.

7. Usando o método ***k-fold cross validation*** obtenha a previsão do atributo **Salario** com um modelo:
   a. K-vizinhos-mais-próximos e o valor de k obtido na alínea anterior
   b. Árvore de regressão
   c. Obtenha a média e o desvio padrão taxa de acerto dos modelos
   d. Verifique se existe diferença significativa no desempenho dos dois melhores modelos obtidos anteriormente (use um nível de significância de 5%). Identifique o modelo que apresenta o melhor desempenho.

5. Referências Bibliográficas

[1]. Christopher Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[2]. Tom Mitchell, Machine Learning. McGraw-Hill, 1997.