Introdução à Aprendizagem Estatística

Projeto de Grupo

Perfil de Especialização em Ciência de Dados – MIEI/MEI/MMC

Docente: Raquel Menezes

2018/2019

A premissa deste projeto é simples. É necessário escolher um conjunto de dados real para o qual se acredite que existem perguntas interessantes para responder. Em seguida, devem experimentar/considerar os vários métodos de aprendizagem estatística, que têm sido abordados nas aulas de "Aprendizagem Automática I", para tentarem encontrar a melhor maneira de responder a essas perguntas. O projeto deve ser executado em grupos, sendo cada grupo composto por 3 (ou 2) elementos.

Documentos a Entregar

O projeto envolve a preparação faseada de 3 documentos, que deverão ser enviados por mail para a docente, *rmenezes@math.uminho.pt*, nomeadamente:

1. A proposta para o projecto – 1 página (**Até 17 nov, sáb**)
   (a) Nomes dos membros
   (b) Descrição do problema
   (c) Descrição do conjunto de dados (dimensões, nomes das variáveis com sua descrição)
   (d) Supervisionado ou não supervisionado?
   (e) Regressão ou classificação?
   (f) Comentários e/ou preocupações?

2. Um *poster* (ou um conjunto de *slides*) para uma apresentação de 5/10 minutos (**Dia 12 dez, 4.f, 14h30-16h30**)
   (a) Descrição dos dados e as perguntas que o grupo está interessado em responder
   (b) Revisão de algumas das abordagens que o grupo experimentou ou pensou experimentar
   (c) Resumo da abordagem final que o grupo usou e porque escolheu essa abordagem
   (d) Resumo dos resultados
   (e) Conclusões

   Ao preparar a apresentação, o grupo deve assumir uma audiência com formação estatística ao nível de regressão linear múltipla, mas não além disso. Pelo que, por exemplo, não deve apenas dizer "Nós fizemos KNN", mas também explicar a ideia básica de como funciona, porque este método pode ser preferível ao da regressão linear, etc.

3. Um relatório do projeto – máximo 10 páginas (**Até 21 dez, 6.f**)

   O relatório deve conter um pouco mais de detalhe sobre o material coberto na apresentação (máximo de 3 páginas). A primeira página deve ser um *resumo executivo* (i.e. versão condensada do documento completo). Em anexo, deve incluir o *poster* ou *slides* da apresentação, assim como deve considerar um anexo técnico com o código R desenvolvido para a realização do trabalho.

   A **avaliação final do trabalho** irá considerar diversos aspetos, nomeadamente a escolha adequada das questões de interesse, a abordagem usada para resolvê-las, o motivo pelo qual você escolheu essa abordagem e as conclusões que o grupo conseguiu extrair.

Possíveis repositórios de dados

1. Dados Abertos Gov.: dados.gov.pt, dados.gov.br, www.data.gov.uk, www.data.gov
2. Kaggel: www.kaggle.com
3. KDD Nugets: http://www.kdnuggets.com/datasets/
4. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
5. StatLib: http://lib.stat.cmu.edu
6. TwitteR: http://cran.r-project.org/web/packages/twitteR/index.html
7. rfigshare: http://figshare.com, http://cran.r-project.org/web/packages/rfigshare/index.html
8. swiss: automaticamente disponível em ambiente R
9. diabetes: disponível em ambiente R debaixo da library "faraway"
10. fat: disponível em ambiente R debaixo da library "faraway"
11. Alguma base de dados (diferente das utilizadas na aula) da library "ISLR"
12. etc