Aprendizagem Automática

Trabalho Laboratorial – grupos de 1, 2 ou 3 alunos

Classificação de Críticas de Cinema do IMDb

1 Dados

A IMDb, a Internet Movie Database, é uma base de dados que consiste em textos de críticas de cinema, recolhidas por Andrew Mass [**?**]. Neste trabalho, os dados são compostos por 50 000 textos de críticas de cinema com as respetivas pontuações, e encontram-se no ficheiro `imdbFull.p`. Este ficheiro contém uma variável do tipo dicionário com dois campos: `data` lista com os textos de criticas de cinema, `target` com a pontuação da crítica (numa escala de 1 a 10). Críticas neutras (pontuações 5 e 6) foram excluídas.
Para mais informação, consultar: `ai.stanford.edu/~amaas/data/sentiment/`

2 Objectivos do trabalho

Este trabalho está dividido em três tarefas:

I. Treinar e avaliar um classificador multi-classe para prever a pontuação da crítica.

II. Treinar e avaliar um regressor para prever a pontuação da crítica.

III. Fazer *clustering* das críticas.

3 Desenvolvimento

Deverá ter em conta os seguintes pontos:

1. **Construção do vocabulário**:
   A construção do vocabulário é um passo fundamental para o bom desempenho dos modelos analisados. Para tal deve proceder à limpeza dos textos e testar os parâmetros que considerar relevantes da função `TfidfVectorizer`. De notar que vocabulários de dimensão elevada podem ter uma influência negativa no desempenho dos modelos.
2. **Classificação e regressão**:
   Use o mesmo vocabulário para estas duas tarefas. Escolha um ou mais modelos de classificação e de regressão. Compare os resultados obtidos na classificação com os obtidos na tarefa de regressão. Para tal quantifique os valores preditos pelo(s) modelo(s)

de regressão de maneira a serem convertidos nas pontuações das críticas (convertidos em valores de 1 a 4 e de 7 a 10).

3. **Metodologias de teste e métricas de desempenho**:
   Escolha a metodologia de teste apropriada de modo a ter uma estimação fidedigna do desempenho dos modelos treinados.

4. ***Clustering***:
   Escolha um ou mais métodos de *clustering* à sua escolha para agrupar críticas de uma forma não supervisionada. Analise os resultados e indique os grupos em que as críticas abordam um tópico específico. Investigue o efeito da variação do número de *clusters* no desempenho dos algoritmos de agrupamento.

4 Ficheiro a Entregar

Deve ser entregue, via Moodle, um único ficheiro Jupyter denominado `AxxxxxAxxxxxAxxxxxTP2.ipynb` (onde `Axxxxx` são os números de alunos do grupo - colocar em ordem crescente). O Jupyter Notebook deve estar devidamente comentado de modo a transmitir claramente os passo dados, a razão para as opções tomadas e uma análise detalhada dos resultados obtidos.