Aprendizagem Automática

Trabalho Laboratorial – grupos de 3 alunos

Classificação de Críticas de Cinema do IMDb

1º Semestre de 2017/2018

Objectivos do trabalho: Este trabalho lida com a análise de críticas de cinema do IMDb, e está dividido em duas principais tarefas: classificação e regressão. Em classificação, o problema consiste em determinar o tipo de crítica baseado no texto do documento. Em regressão o objetivo é prever a pontuação da crítica, numa escala de 1 a 10. O desempenho dos modelos projetados está intrinsecamente ligado à construção do vocabulário na representação tf-idf, e por isso, este é também um tema de análise deste trabalho prático.

1. **Classificação:**
   Deverá desenvolver e testar modelos para dois tipos de problemas de classificação: binária e multi-classe. Em classificação binária o problema consiste em determinar se uma dada crítica é positiva ou negativa. Em multi-classe, objetivo é prever a pontuação da crítica, e para tal considera-se que existem 8 classes, compostas pelas criticas de 1-4 e de 7-10.
   Use em ambos os problemas um discriminante logístico. Deverá igualmente testar outro modelo de classificação à sua escolha. Os resultados deverão ser comparados com os obtidos com modelos de regressão.
2. **Regressão:**
   Deverá projetar um modelo de regressão linear para prever a pontuação de uma crítica num escala 1 a 10 valores baseado no texto da mesma. Os resultados obtidos devem ser comparados com os da classificação multi-classe e binária. Note que para comparar os resultados do regressor com os da classificação binária, deverá previamente converter as pontuações estimadas em positivas ou negativas.
3. **Representação tf-idf:**
   Deverá ter em conta os seguintes pontos:
   - Estudar se a inclusão de n-gramas é benéfico para o desempenho dos modelos projetados.
   - Testar se stemming dos documentos antes da representação tf-idf é benéfico para o desempenho dos modelos projetados.
   - Deverá analisar como o desempenho dos modelos projetados é afetado pela dimensão do vocabulário. Deverá igualmente averiguar qual a dimensão mínima do vocabulário para a qual o desempenho dos classificadores binários seja ainda próximo dos melhores resultados obtidos.

Dados: A IMDb, a Internet Movie Database, é uma base de dados que consiste em textos de críticas de cinema, recolhidas por Andrew Mass [1], e que se encontra disponível em `ai.stanford.edu/ amaas/data/sentiment/`. A ficheiros nesta base de dados encontram-se

guardados em duas diretorias de topo, `train/` com os dados de treino, e `test/`, com os dados de teste. Por sua vez, em cada uma destas diretorias encontram-se duas sub-diretorias `pos/` com os exemplos positivos e `neg/` com os negativos. A divisão das críticas em duas classes, positivas e negativas, é para a tarefa de classificação. Críticas positivas têm uma pontuação superior ou igual a 7 (numa escala de 1 a 10) e as negativas uma pontuação inferior ou igual a 4. Críticas neutras (pontuações 5 e 6) foram excluídas. Para a tarefa de regressão, a informação sobre o valor da pontuação (rating) encontra-se no nome do próprio ficheiro. Na tarefa de regressão é necessário extrair esta informação dos nomes dos ficheiros para criar a "matriz" $\mathbf{Y}$ de $1 \times N$, com as saídas desejadas (onde N é o número de documentos).
(Para mais informação sobre esta base de dados, ler o ficheiro `README` disponibilizado com a mesma, e para carregar a base de dados em ambiente Python, consultar os acetatos da disciplina sobre esta matéria).

Metodologias de Teste: O desempenho dos algoritmos projetados deve ser avaliado com base nos resultados obtidos no conjunto de teste. Para as tarefas de classificação deve reportar a probabilidade de erro (ou acerto) e a matriz de confusão, mas também deve usar outras métricas de desempenho utilizadas em problemas de classificação binária. Para modelos de regressão, a métrica de desempenho é o coeficiente de determinação R^2.

Etapas do Trabalho: Para diferentes etapas do trabalho enumeradas no início do enunciado, tenha em conta o seguinte:

- Os documentos de texto são representados numericamente com o modelo tf-idf. Deve descrever em detalhe todo processo de limpeza dos documentos bem como bem como a ordem dos n-gramas tido em consideração no cálculo da matriz documento-termo.
- Nas tarefas de classificação, testar o desempenho dos discriminantes logísticos para diferentes dimensões do vocabulário. Testar igualmente se a regularização dos coeficientes é benéfico para o desempenho dos discriminantes. Converter os resultados do classificador multi-classe para binário (classes positivas e negativas) e comparar com os resultados obtidos com um classificador binário. Além dos discriminantes logísticos, escolha outro modelo de classificação, indicando a razão da escolha, e compare os resultados.
- Deverá testar modelos de regressão linear sem regularização e com regularização *ridge* e *lasso*. Testar diferente dimensões do vocabulário. Compara os resultados obtidos com os dos classificadores.
- Aplicar o método de regularização *lasso* para selecionar vocabulários de diferentes dimensões, e testar o desempenho do discriminante logístico no problema de classificação binária. Testar igualmente este problema com outro classificador e com os mesmos vocabulários usados para o discriminante logístico. Repetir os testes com vocabulários das mesmas dimensões, mas obtidos definindo o parâmetro `max_features` na função `TfidfVectorizer`.

Elaboração do Relatório:

- O relatório terá no máximo 8 páginas e deverá ser formatado de acordo com o modelo "ACM_Large" da revista ACM - Association for Computer Machinery (ver página http://www.acm.org/publications/authors/submissions). Os membros do grupo devem estar claramente identificados com número e nome no campo referente aos autores, e o relatório deverá ser bem estruturado: ter uma introdução, uma conclusão, descrição dos métodos usados, das experiências efetuadas, resultados obtidos, referências usadas, etc.
- Não inclua no relatório o código implementado.
- Deve comentar os resultados obtidos e as possíveis causas para os (bons/maus) desempenhos, e quando achar pertinente, complementar o seu raciocínio com gráficos ou imagens.
- Inclua na bibliografia todo o material consultado para elaborar o relatório.
- IMPORTANTE: Entregar unicamente o ficheiro do relatório. Terá que ser um ficheiro ".pdf" com o nome: `TP2_Axxxx_Axxxx_Axxxx.pdf` onde os "`Axxxx`" correspondem aos números de aluno dos 3 membros do grupo. Colocar os número em ordem crescente.

Referências

[1] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.