INF-0615– Aprendizado de Máquina Supervisionado I

Trabalho 01

Predição da Popularidade de Notícias em Redes Sociais

Data de Entrega: 04/09/2022

1 Descrição do Dataset

Um dos maiores interesses das empresas, empreendedores e meios de comunicação é prever a popularidade de um anúncio ou de algum informação divulgada nas redes sociais. Um maior número de compartilhamentos significa um maior alcance dos conteúdos indicando maior número de pessoas interessadas. Neste contexto, o objetivo desse trabalho é predizer o quão popular uma informação será na rede social por meio da predição do número de compartilhamentos de um *post*. Para isso, um conjunto de atributos em relação às características do *post* são dados abaixo:

- **n_tokens_title**: Número de palavras no título do *post*
- **average_token_length**: Tamanho médio das palavras no conteúdo do *post*
- **num_keywords**: Número de palavras-chave nos meta-dados
- **kw_avg_max**: Número médio de compartilhamentos da melhor palavra-chave do *post*
- **global_subjectivity**: Mensuração da subjetividade do texto
- **global_sentiment_polarity**: Mensuração da polaridade sentimental do texto
- **global_rate_positive_words**: Taxa de palavras "positivas" no conteúdo do *post*
- **global_rate_negative_words**: Taxa de palavras "negativas" no conteúdo do *post*
- **rate_positive_words**: Taxa de palavras "positivas" apenas considerando as palavras não-neutras
- **rate_negative_words**: Taxa de palavras "negativas" apenas considerando as palavras não-neutras
- **avg_positive_polarity**: Polaridade média das palavras "positivas"
- **avg_negative_polarity**: Polaridade média das palavras "negativas"
- **log_n_tokens_content**: Logaritmo do número de palavras no conteúdo do *post*
- **log_num_hrefs**: Logaritmo do número de links no *post*
- **root2_num_self_hrefs**: Raiz quadrada do número de links para outros *posts* na mesma rede social
- **log_self_reference_max_shares**: Logaritmo do maior número de compartilhamentos de um artigo referenciado no *post* na mesma rede social
- **log_self_reference_avg_sharess**: Logaritmo do número médio de compartilhamentos de um artigo referenciado no *post* na mesma rede social
- **weekday**: Dia da semana em que o *post* foi (ou será) publicado
- **target: Número de compartilhamento que o *post* terá nas redes sociais (esse é o valor alvo que vocês devem predizer)**

2 Tarefas

Pedimos que vocês:

1. Inspecionem os dados. Quantos exemplos vocês tem? Como vocês irão lidar com as features (atributos) categóricas, se houverem? Há exemplos com features sem anotações? Como vocês lidariam com isso?
2. Apliquem alguma técnica de normalização de forma a deixa os dados mais bem preparados para o treinamento (Min-Max, Z-Norma, etc).
3. Como *baseline*, treinem uma regressão linear utilizando todas as features para predizer o número de compartilhamentos de um *post* em uma rede social. Reportem o erro nos conjuntos de treinamento, validação e teste.
4. Implementem soluções alternativas baseadas em regressão linear através da combinação das features existentes para melhorar o resultado do *baseline*. Comparem suas soluções reportando os erros no conjunto validação. Tomem **apenas a melhor solução baseada no conjunto de validação** e reportem o erro no conjunto de teste. **Lembrem-se que as features categóricas podem ser incluída no modelo mas fora do processo de combinação. Vejam o Ex01.R como referência**.
5. Implementem soluções alternativas baseadas em regressão linear aumentando os graus das features (regressão com polinômios) para melhorar o resultado obtido no *baseline*. Plotem o erro no conjunto de treinamento e validação pelo grau do polinômio. Identifiquem as regiões de *underfitting*, ponto ótimo e *overfitting*. Tomem **apenas o melhor modelo polinomial baseado no conjunto de validação** e reportem seu erro no conjunto de teste.
6. Escrevam um relatório de no máximo 5 páginas:
   (a) Descrevam o que foi feito, bem como as diferenças entre o seu melhor modelo e o seu baseline;
   (b) Reportem o erro do melhor modelo de todos no conjunto de teste. Lembrem-se que o melhor modelo de todos deve ser escolhido baseado no erro no conjunto de validação.
   (c) Uma Seção de conclusão do relatório explicando a diferença entre os modelos e o porquê que estas diferenças levaram a resultados piores ou melhores.

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *T01_train_set.csv*: conjunto de dados para treinamento;
- *T01_val_set.csv*: conjunto de dados para validação;
- *T01_test_set.csv*: conjunto de dados de teste retido pelo professor (**será disponibilizado na quinta-feira anterior ao prazo final da submissão**).
- *T01_codigo_de_apoio.R*: código de apoio a partir do qual vocês devem realizar o trabalho.

4 Avaliação

O dataset foi previamente dividido aleatoriamente em três conjuntos — treino, validação e teste — e apenas os dois primeiros serão disponibilizados para que vocês implementem as suas soluções.

Na quinta-feira anterior ao prazo final de submissão, iremos disponibilizar no Moodle o conjunto de teste e iremos avisá-los pelo canal da disciplina no Slack. No relatório, vocês devem reportar tudo que foi pedido na seção Tarefas.

A avaliação consistirá da análise do relatório e do código submetidos no Moodle. Iremos avaliar se as tarefas pedidas foram realizadas, como o treinamento e validação foram feitos, os resultados reportados e as conclusões.

Observações sobre a avaliação:

- O trabalho deverá ser feito em duplas ou trios, podendo haver repetição dos membros a cada trabalho;
- O código (arquivo .R) e o relatório (formato .pdf) deverão ser submetidos no Moodle por **apenas um integrante do grupo**;
- Não se esqueçam de listar os nomes dos integrantes do grupo no início do relatório e no código;
- As notas do trabalho serão divulgadas em até uma semana após o prazo da submissão;