EP 2: Transformers, Regressão e Quantização

Entrega: Domingo, 10/12/2023 (improrrogável)

Nota: Se você possui uma outra proposta de trabalho de final de curso que potencialmente esteja relacionado ao seu trabalho de pesquisa de pós-graduação, por favor, apresente uma proposta que nós podemos discutir como transformá-la no enunciado para este trabalho final.

1 Motivação

O objetivo deste exercício é aumentar a familiarização dos alunos com o refinamento de encoders na arquitetura Transformers (BERT-like). Também iremos verificar o uso desta arquitetura no âmbito da regressão numérica. As redes neurais podem ser usadas tanto para regressão quanto para classificação, porém, na arquitetura Transformers (BERT-like) elas são treinadas totalmente na tarefa de classificação. Ou seja, as tarefas de pré-treinamento na previsão de palavras mascaradas são tarefas de classificação de vetores em um número muito grande de classes, que correspondem às palavras. Por outro lado, tarefas de detecção da ordem da sentenças são classificações binárias.

Sendo assim, não é de se espantar que redes pré-treinadas para classificação não obtenham resultados muito bons em tarefas de regressão. O objetivo desse trabalho é chamar a atenção para este fato e para propor possíveis soluções. Uma forma de amenizar o problema de baixa performance na regressão é a quantização, ou seja, transformar um contínuo de valores em faixas de tamanho possivelmente variado, de forma a gerar um número pequeno de categorias, transformando um problema de regressão no problema de classificação.

> ***Importante:*** *Em todas as tarefas a seguir, vocês devem usar o refinamento (finetuning) da rede neural BERTimbau, obtida a partir do site Huggingface.*

1.1 Dados

Nós utilizaremos o corpus de avaliações da B2W, e estaremos interessados em apenas uma coluna: `review_text`, que será a referida daqui para frente como `texto`. As demais informações deverã ser computadas no pré-processamento do córpus.

Você pode utilizar um sub-córpus de no mínimo 10 mil sentenças, ou utilizar todas as sentenças disponíveis, que deverá ser particionado em três córpus de treinamento, validação e teste, nas proporções usuais que você dará deverá mencionar no seu artigo de descrição.

2 Tarefa 1: Regressão de densidade de vogais

A tarefa de regressão que iremos prever com redes neurais Transformers será a densidade de vogais em um trecho de texto (DV). Esta grandeza numérica é obtida considerando-se apenas os caracteres que correspondem às letras do alfabeto português, maiúsculas ou minúsculas [A-Z,a-z,Ç,ç,Ã,ã,...], contando o número de vogais com ou sem acento, e dividido pelo número total de letras. Desta forma, estaremos ignorando todos os caracteres de espaço e pontuação, assim como os dígitos e outros caracteres não-letra.

Por exemplo, na sentença obtida diretamente do córpus B2W-reviews:

> Meu filho amou! Parece de verdade com tantos detalhes que tem!

Esta sentença possui 63 caracteres, porém descontando os espaços em branco e os sinais de pontuação ela possui apenas 50 letras, das quais apenas 23 são vogais. Sendo assim a densidade de vogais da sentença é de 0,43.

É importante notar que quando se faz a tokenização seguida do embedding, os caracteres que compõem a palavra não estão mais presentes, então a densidade de vogais é um número totalmente abstrato, independente da língua, e do estilo de texto.

No pré-processamento você deverá calcular a densidade de vogais de cada uma das sentenças do seu córpus, tanto na parte de treinamento, quanto na parte de validação, quanto na parte de teste.

2.1 Loss

Apenas a título de sugestão, propomos que a função de loss a ser usada durante o treinamento seja a função MSE (erro quadrático médio). Outras medidas podem ser usadas, mas lembre-se que a medida de entropia cruzada só deve ser usada para tarefas de classificação.

Mencione qual função foi utilizada na descrição do experimento no seu artigo.

2.2 Avaliação

O seu experimento deverá ser avaliado por diversas métricas. Estamos solicitando que sejam utilizadas no mínimo as seguintes métricas de avaliação

- raiz quadrada do erro quadrático médio (RMSE)
- erro absoluto médio (MAE)
- erro proporcional absoluto médio (MAPE)
- R^2
- correlação de Pearson

Você deverá comparar seus resultados com três baselines: um valor fixo correspondente à densidade de vogais de todo o seu córpus, a densidade de vogais da primeira palavra da sentença e a densidade de vogais da última palavra da sentença, tá e calcular as estatísticas acima para estas três baselines.

3 Tarefas 2 e 3: Quantização

Vamos realizar duas tarefas de classificação quantizada, em que teremos três categorias para classificar a densidade de vogais de uma dada sentença. Na Tarefa 2, vamos classificar uma sentença de acordo com três classes. Na classe 1, estão as sentenças com densidades inferiores a $\frac{1}{3}$; na classe 2 estão as sentenças com densidades entre $\frac{1}{3}$ e $\frac{2}{3}$; e na classe 3 estão as sentenças com densidades superiores a $\frac{2}{3}$.

Note que na tarefa 2 as classes não estão balanceadas, e é muito provável que classe 2 contenha quase a totalidade da sentenças do córpus.

Na Tarefa 3 vamos criar três classes balanceadas, ou seja, o número de sentenças de cada classe deve ser um terço da sentenças do córpus.

Em ambas as tarefas você deve medir a acurácia total e a acurácia de cada uma das classes, assim como deve medir a sensibilidade e a especificidade de cada uma das três classes.

Não há necessidade de criar baselines para estas tarefas, apenas comparar as avaliações de uma com a outra.

Instruções entrega

Entregar um zip no e-Disciplinas (moodle) contendo os seguintes elementos:

1. Um diretório *src* com os seus arquivos em python3.
2. Um diretório *data* com os arquivos com os córpus usados.
3. Um arquivo *README* com as instruções de pré-processamento, treinamento e teste
4. Um arquivo *relatorio-ep2.pdf* com os seguintes conteúdos:
   (a) Uma descrição das tarefas realizadas, e das medidas de avaliação.
   (b) Uma descrição dos experimentos das configurações e dos híper-parâmetros utilizados.
   (c) Duas tabelas, uma com os resultados da Tarefa 1 e de seus baselines, e outra comparando os resultados das Tarefas 2 e 3.
   (d) Uma discussão dos seus resultados.