Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) – Campus Maracanaú

Redes Neurais (2020.2)

Professor: Amauri Holanda (`amauri01 at gmail.com`)

Classificação de texto — Entrega: 26/03

1 Trabalho computacional

Neste trabalho computacional você empregará **redes recorrentes** a um problema real da área de aprendizagem de máquina: análise de sentimento através de texto. O problema consiste em definir se um determinado texto corresponde a um comentário positivo, negativo ou neutro. Análise de sentimento pode ser aplicada em vários tipos de texto, incluindo *reviews* de filmes, *tweets*, hospedagem. Conjunto de dados para problemas de análise de sentimento pode ser encontrados facilmente na internet.

Recomenda-se utilizar uma representação de texto na qual cada palavra é representada por um vetor *one-hot* — ou seja, somente um valor igual a 1 e os outros valores iguais a 0. Para isso, você deve primeiro criar um vocabulário de palavras. Cada palavra consistirá de um vetor *one-hot*, do tamanho do vocabulário, em que o índice do valor 1 corresponde ao índice da palavra no vocabulário. Note que essa abordagem já foi adotada na implementação de rede recorrente vista em sala.

Nesse trabalho, você deve:

- Treinar uma rede recorrente (e.g., Elman);
- Plotar a evolução da função custo (loss) ao longo do treinamento (épocas);
- Reportar taxas de acerto (ou erro) nos conjuntos de treinamento e teste.

1.1 Dataset

A seguir está uma descrição *adaptada* do conjunto de dados utilizados neste trabalho e disponível em `https://www.kaggle.com/ranjitha1/hotel-reviews-city-chennai`.

Os dados foram obtidos de Trivago – Índia. Um `script python` foi executado para examinar os pedidos de obtenção e fazer esses pedidos de forma explícita para obter os dados necessários no JSON. Estes dados foram analisados e escritos em um arquivo `.csv`.

Os dados estão na forma de um arquivo `.csv` com mais de 4000 comentários. Existem 5 colunas:

1. **Coluna 1**: Nome do hotel;
2. **Coluna 2**: Título do *review*;

3. **Coluna 3**: Texto da *review*;
4. **Coluna 4**: Sentimento da *review* (1: Negativo 2: Média 3: Positivo)[1];
5. **Coluna 5**: Percentual de classificação.

Este conjunto de dados consiste em reviews de pessoas reais. Então, esse conjunto trará uma experiência real sobre como lidar com dados de texto.

1.2 Entrega

Um relatório deve ser enviado até a data especificada no início deste documento via *google classroom*. Preferencialmente, o relatório deve consistir de *um único python notebook* com as respostas, gráficos, comentários e códigos.

Cada relatório pode ser desenvolvido por até duas pessoas.

2 Material de apoio

1. **Código: Tratamento de dados**
   Foram elaborados alguns *snippets* de código na linguagem de programação `python` referentes a possíveis passos necessários para a realização do trabalho. Estão acessíveis a partir do endereço
   `https://github.com/IFCE-Mestado-Ciencia-da-Computacao/PGM-2017.2/blob/master/Lista-2.ipynb`;

Bons estudos!

[1]Existem três valores de sentimento. 1 representa uma revisão negativa, enquanto 3 representa uma positiva.