SEASONS OF CODE 2024

Sentiment Analysis and Text Generation using Many-to-One LSTMs

Mentors: Shreyas Katdare & Dion Reji

Date: May 31, 2024

Week 1 - Assignments

We'll kick things off with Python and some essential packages. If you are familiar with these, chill! If not, here's how to jumpstart your journey:

1. A detailed Tutorial (pdf) for python can be found in the files section of the GitHub repo. Refer to chapters 3,4,5,7, and 9 of the document to learn the basics of Python. (If you already know Python, you may use these to review/recall your Python knowledge). If you need a You-tube tutorial, Click here.

2. `numpy` is a powerful and useful package extensively used in Machine Learning. You can find a detailed `numpy` tutorial (pdf) in the files section of the repo. Refer to Chapter 3 to get started with `numpy`. For a You-tube `numpy` tutorial, Click here.

3. `pandas` is another useful package used for reading, and editing data. Please refer to this website to learn about `pandas`. A useful You-Tube tutorial can be found here.

4. We will be working with *Jupyter Notebook*, for this project. Just like any project/coding-assignment, we need a code-editor to work with. So, to work with Jupyter Notebook in VS code, please refer to the You-Tube video here. You may use any other code-editors, however we expect your final submission as Jupyter Notebook.

After being familiarized with basic python, `numpy`, and `pandas`, try out the following problem. **NOTE that this assignment is meant to be submitted.** Also, this assignment doesn't contribute to your final project. This is meant to make sure that you are familiar with the basics, and to help you review your understandings and clear your doubts.

Problem Statement

In this assignment, you will be working on a regression problem. Regression is a type of predictive modeling technique that estimates the relationship between a dependent (target) and independent (predictor) variables. It is widely used in statistics, machine learning, and data science for tasks such as forecasting and determining relationships between variables.

You will implement linear regression from scratch using gradient descent.

Linear Regression

Linear regression is a method for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted x.

The linear regression model can be written as:

$$y = h_\theta(x) = \theta_0 + \theta_1 x$$

where:

- y is the dependent variable.
- x is a vector of independent variables.
- $\theta = (\theta_0, \theta_1)$ is a vector of coefficients.

Our task is to find the vector θ of co-efficients. For that, we define an Error Function, and seek to minimize it.

Error Function (Loss Function)

The cost function $J(\theta)$ for linear regression is defined as:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

where m is the number of training examples.

The gradient of the cost function with respect to θ is:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

where $x_j^{(i)}$ is the j-th feature of the i-th training example.

In vectorized form, this can be written as:

$$\nabla J(\theta) = \frac{1}{m} X^T (X\theta - y)$$

In order to find the optimal θ, we minimize $J(\theta)$ using an algorithm called Gradient Descent

Gradient Descent

The Gradient Descent algorithm is in general used for minimizing any error function $J(\theta)$. It is given in Algorithm 1.

Algorithm 1 Gradient Descent

```
Initialize θ with some value
Set learning rate α
Set max number of iterations
for i = 1 to max number of iterations do
    Compute the gradient ∇J(θ) using the current θ
    Update θ using the formula: θ = θ − α · ∇J(θ)
end for
return θ
```

Resources

You can read more about these things as indicated below:

- Linear Regression - Wikipedia
- Gradient Descent - Wikipedia

Tasks to be completed

You are provided with a `Regression.ipynb` file. (See inside the Week 1 folder of the GitHub repo). We have partially done the work for you. Your task is to complete the TO-DO's mentioned inside the code.

- TODO 1: Read data from `data.csv` file provided to you. (See inside the Week 1 folder of the GitHub repo), using `pandas`
- TODO 2: Complete the `normalize` function.
- TODO 3: Complete the function `compute_cost` to calculate the error function for linear regression
- TODO 4: Implement the gradient descent for linear regression. Your only task here is to update theta properly by finding the gradient correctly.
- TODO 5: Find the optimal value of θ and cost history by calling the `gradient_descent` function. Also, print them.

Submission

You are supposed to submit the Jupyter Notebook file after completion.

The deadline for submission is Monday 03/06/2024 EOD. Further informations regarding the submission will be conveyed later.