Crab Age Prediction

Background

Crab is very tasty and many countries of the world import huge amount of crabs for consumption every year. The main benefits of crab farming are, labor cost is very low, production cost is comparatively lower and they grow very fast. Commercial crab farming business is developing the lifestyle of the people of coastal areas. By proper care and management we can earn more from crab farming business than shrimp farming. You can raise mud crabs in two systems. Grow out farming and fattening systems.

Dataset

8 features: sex, length, diameter, height, weight, Shucked Weight, Viscera Weight, and Shell Weight. 1 output: age. Size: 2893 instances

Task 1: Implement stochastic gradient decent (SGD), batch gradient decent (BGD), and normal equation to learn linear regression model. Use 90 percent (row 1- row 2603) as your training data and the last 10% as your testing data. For the category type of the first feature, you can set F (Female) as 0, M (Male) as 1, and I (indeterminate) as 2.

1) Show the plot of $J(\theta)$ for the iterative procedure of gradient decent for SGD and BGD;
2) Report the mean square error (MSE) for all methods in the testing set.

$$MSE = \frac{1}{N_{test}} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

where N_{test} is the testing dataset's size, $\hat{y}^{(i)}$ is the predicted value, and $y^{(i)}$ is the true value.

Task 2:

1) Try to remove the first feature;
2) Try to do normalization for each feature by $\frac{x_j - \mu(x_j)}{\sigma(x_j)}$, where $\mu(x_j)$ is the mean value of the feature x_j for the whole dataset, $\sigma(x_j)$ is the standard deviation of the feature x_j for the whole dataset.

Do they help to improve the MSE? Briefly explain why or why not.

Task 3: Implement locally weighted regression (LWR) using gradient descent (SGD or BGD) and normal equation, report the MSE in the testing data.

1. **We run plagiarism check for submitted code. Please don't look for solutions online.**
2. It's up to you how much training data you will use to train your regression model. To make your training faster, you can use part of it.
3. You will **not receive any credit if you directly use off-the-shelf machine learning tools** for SGD, BGD, and normal equation. Mathematic computation tools such as Numpy and other basic tools, e.g., matplotlib are allowed to use.
4. Submit your code **(.py file)** and your report (**no more than 1 page**, no format requirement, **.pdf file**). Not following the submission requirement, e.g., **code in txt/pdf file, exceeding report page limit, etc. will receive a credit penalty.**