Homework 1
COMP9417, Machine Learning and Data Mining
T3, 2019

Introduction

In this homework, you work on a learning problem where you have to implement a linear regression model and evaluate it.

You will use a publicly available dataset "Advertising Data" which consists of the sales of a product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

We would like to predict the sale from the budgets spent on TV advertisements first and then repeat the regression for predicting the sale from radio and then newspaper advertisements.

You can start by downloading the dataset "Advertising.csv".

1. Pre-processing:

One important per-processing step in most machine learning problems is feature normalisation. Feature normalisation is rescaling the features such that they all have similar scales. This is also important for algorithms like *Gradient Descent* to ensure the convergence of the algorithm.

One of the common normalisation techniques is called min-max normalisation, where each feature is scaled to range between [0,1]. In this normalisation, for each feature, you have to find the minimum and maximum value in all your samples and then use the following formula to make the transformation:

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

After applying this normalisation, the minimum value of your feature will be 0 and the maximum value will be 1.

So, in the first step of this homework, you can start by creating a feature vector which includes the TV, radio and newspaper budget which are the features we will use to predict the sale and then apply min-max normalisation to your features. You can test whether you did the normalisation correctly or not by checking the minimum and maximum value for each of your features.

2. Creating test and training set

In this step, you have to create training and test sets. Please use the first 190 rows of the data as training set and keep the 10 remaining one (from 191 to 200) as test set which we will use later to evaluate the regression model.

3. Gradient descent

Now in this part, you need to fit a regression model that predicts the sale from the budget spend on TV advertisement; so, you have to estimate the regression parameters θ from the training set.

The main objective of linear regression is to minimize the cost function $J(\theta)$:

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

Where in this homework:

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1$$

such that $x_0 = 1$, and x_1 corresponds to TV advertisement budget feature.

In batch gradient descent, you can update the parameters iteratively using the following update rule:

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

Please write a piece of code to estimate parameters θ for the advertising problem.

You can set the initial value of your parameters as ($\theta_0 = -1, \theta_1 = -0.5$) and also use the learning rate of $\alpha = 0.01$ and maximum iteration of 500.

4. Visualization

You can visualize the changes in you cost function $J(\theta)$, at each iteration. You just need to compute the value for your cost function at each step using the value of your parameters at that step and then plot your cost function over iteration steps.

5. Evaluation

Now, it is time to evaluate your estimated regression model on the training and test data using one of the evaluation metrics. Here, you can use Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2}$$

Compute the RMSE once for the training set and once for the test set to see if your model generalises well on unseen samples or not.

6. Repeating for the other two features

Now, in this part of exercise, you want to compare your model with other models that use the radio advertising budget or newspaper advertising budget. In this step, you just need to repeat the step 3 (Gradient descent) to find the parameters for predicting the sale once using only radio feature and once using only newspaper feature.

Now evaluate these two new models on the test set and compare your three regression models to see which one gives the best prediction on your test set.

Due date: Friday 11.10.1019 by 5:00pm

What to submit: (5 marks)

You can report the following in a .pdf file:

1. You have to report the θ parameters in step 3 when you are using TV feature. (2 marks)
2. A plot, which visualises the change in cost function $J(\theta)$ at each iteration. (1 mark)
3. RMSE for your training set when you use TV feature. (0.5 mark)
4. RMSE for test set, when you use TV feature. (0.5 mark)
5. RMSE for test set, when you use Radio feature. (0.25 mark)
6. RMSE for test set, when you use newspaper feature. (0.25 mark)
7. Compare the performance of your three models and rank them accordingly. (0.5 mark)

You are also required to add (copy-paste) your code(s) at the end of your report.