Homework 1
COMP9417, Machine Learning and Data Mining
T1, 2020

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 "Real estate", adapted for this homework exercise as "house_prices.csv". This dataset contains 400 real estate records, each consisting of the house age, distance to the nearest MRT station, number of convenience stores in this location, and the house price of unit area.

We would like to predict the price from the house age first and then repeat the regression for predicting the price from the distance to the station and then number of convenience stores.

You can start by downloading the dataset "house_prices.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 house age, distance to the station and number of stores, which are the features we will use to predict the unit house price, 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 300 rows of the data as training set and keep the 100 remaining one (from 301 to 400) as test set which you will use later to evaluate the regression model.

3. Stochastic gradient descent

Now in this part, you need to fit a regression model that predicts the unit house price from the house age; 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 the house age feature.

In **stochastic** gradient descent, you can update the θ_0 and θ_1 parameters iteratively using the update rule presented on slide 40 in Week1 Lecture (Week1_Regression_Part1.pdf)

Please write a piece of code to estimate parameters θ for the real estate 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 iterations of 50.

4. Visualization

You need to 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 distance to the station and the number of stores nearby. In this step, you just need to repeat the step 3 (Stochastic Gradient descent) to find the parameters for predicting the price once using only **distance** feature and once using only **number of stores** 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: Monday 16 March 2020 by 5:00pm

What to submit: (5 marks)

You need to report the following in a **PDF** file:

1. The θ parameters (θ_0, θ_1) from step 3 when you are using house age 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 house age feature. (0.5 mark)
4. RMSE for test set, when you use house age feature. (0.5 mark)
5. RMSE for test set, when you use distance to the station feature. (0.25 mark)
6. RMSE for test set, when you use number of stores feature. (0.25 mark)
7. Compare the performance of your three models and rank them accordingly. (0.5 mark)

!!!!You are required to copy and paste all your code at the end of your report, AND also to submit the Python file(s) with all your code as a separate .zip file!!!!