Data Driven Optimization
Tutorial 1

A.K. Cherukuri

April 25, 2023

1 Linear Regression

Go to Brightspace Content Tutorials Tutorial 1 to find the startup data.csv dataset. This dataset contains records of several startups and has logged how much each start up spent on their RD, Administration and Marketing, how much Profit they made, and in which city they're located (either 0 or 1). For this assignment, we want to predict the company's profit, based on what they spent on their RD, Administration and Marketing.

1.1 Simple Linear Regression

First, we'll predict the company's profit solely based on their RD expenses. Steps:

1. Load the dataset to a matrix in Matlab using

   `csvread('startup_data.csv',1,1).`

2. Define the independent (input) and dependent (output) variables and store them under x and y. Use the first 100 entries of the dataset.

3. Split the dataset into a train and test set, using a 80/20 % split. To do this, you can use the

   `cvpartition()`

   function.

4. Create x_{train}, y_{train}, x_{test} and y_{test}.

5. Calculate the optimal values for θ using x_{train} and y_{train} How many θ values do you think you'll get?

6. Create a scatter plot of x_{train} and y_{train}, and plot your fitted line $y_{\text{pred}} = \theta^{\top} x$.

7. Now, see how well your model fits the test set. Repeat the same steps for x_{test} and y_{test}.

8. Determine the error on the test set and create a histogram plot of the error using the function

`plotterrhist()`.

1.2 Multiple Linear Regression

We'll now add more input data to the model, by including the Administration and Marketing expenses. Repeat the steps as above, excluding the plots.

We have only used 100 rows of the data set so far. Try to run your script for 100,1000 and 10000 data points and use

```
tic,
toc,
```

to log the time it takes to find the fit. Do you see a potential problem arising?

2 Logistic Regression

For this assignment, we'll predict whether a person has made an purchase from an social media advertisement, based on their age and estimated salary. Go to Brightspace Content Tutorials Tutorial 1 to find the

`advertisement.xlsx`

dataset. We can follow partly the same steps as for linear regression, however, we have to solve this using an iterative method, instead of solving it directly. The steps required to solve the maximum likelihood problem are described in the lecture slides.

1. Load the dataset in Matlab using

   `xlsread()`.

2. Define the independent (input) and dependent (output) variables and store them under x and y.

3. Normalize x, to reshape the data which improves numerical conditioning, using the function

   `normalize()`.

4. Split the dataset into a train and test set, using a 80/20 % split, to create x_{train}, y_{train}, x_{test} and y_{test}.

5. We will solve this problem using gradient ascent, since we're dealing with *maximizing* the likelihood. For each iteration, compute the log likelihood, the gradient and update θ.

6. Once you have obtained θ^*, compute $y_{\text{pred}} = \theta_*^T x$ and determine which entries belong to 0 or 1.

7. Define the decision boundary

8. Make a scatter plot of the predictions and the decision boundary. For the plot, x will be the age and y will be the estimated salary.