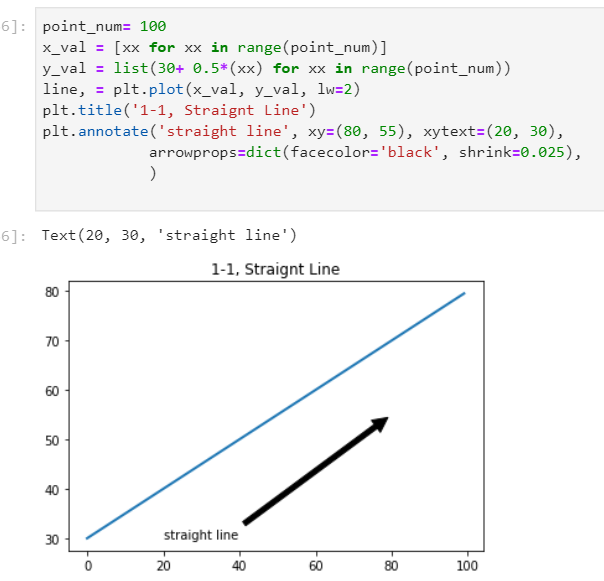
CS-596 Machine Learning Homework Assignment 1

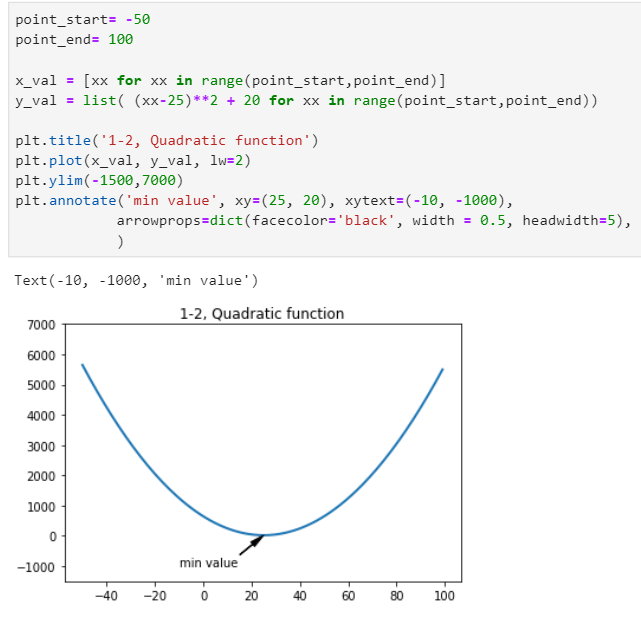
Hsuan Yu, Liu Red ID:823327369

**Question 1**

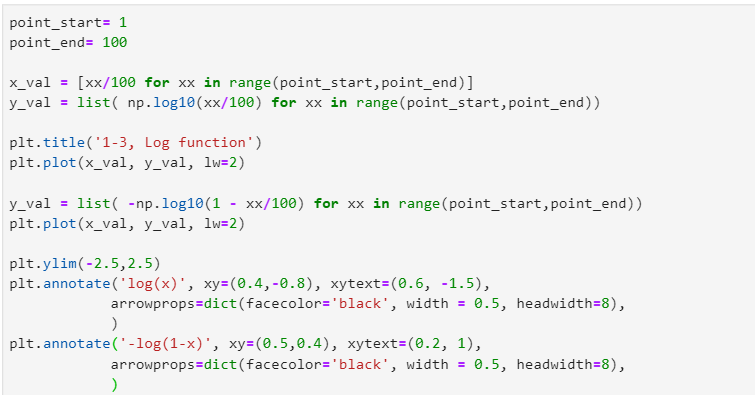
* Straight line where

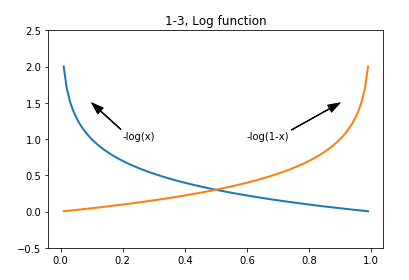


* Quadratic function:

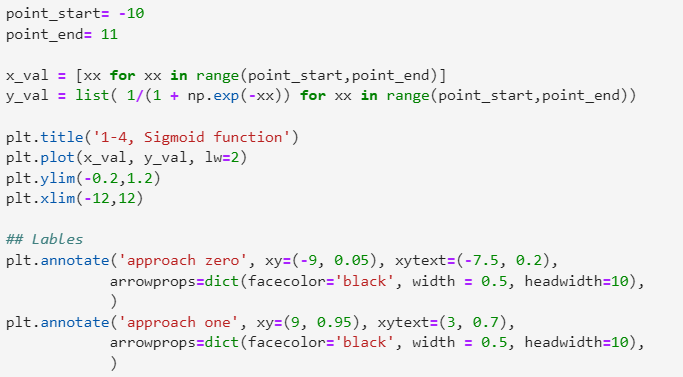


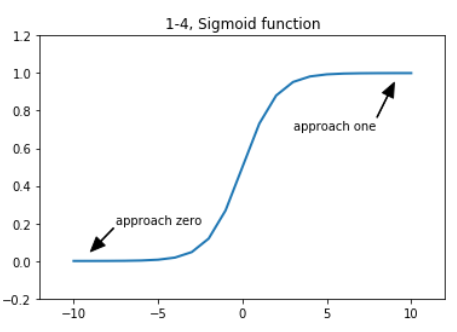
* Log function, and ,





* Sigmoid function,





**Question 2**

You are given the following tasks, all of which can be solved with a certain type of machine learning algorithms.

1. Classify a house to be single family or townhouse. A training set is available. Each training sample is provided with multiple features, including number of bed rooms, number of bathrooms, and house type (single family or townhouse).
2. Classify an email to be spam or not. Users already identified some emails as spam ones.
3. Human tumor Microarray data are provided as a matrix where rows correspond to genes and columns to tissue samples. The task is to cluster columns (or samples) to identify disease profiles: tissues with similar disease should yield similar expression profiles.

Which statement is correct? (Mark one)

1. i) unsupervised learning with discrete predictions;

ii) supervised learning with continuous predictions;

iii) supervised learning with continuous predictions;

1. i) supervised learning with discrete predictions;

ii) supervised learning with discrete predictions;

iii) unsupervised learning with discrete results;

1. i) supervised learning with discrete predictions;

ii) supervised learning with continuous predictions;

iii) unsupervised learning with discrete results;

1. All the three scenarios can be solved by unsupervised learning.

**Question 3:**

**Task:**

* Goal: Predict a rating of business based on its review. Usually, people give high score with positive words. Thus, it might be able to predict a score based on a review.
* Input: A table where rows correspond to reviews and columns to numbers of words which exist in reviews.
* Output: A star between 0 to 5.

**Data preparation:**

On the famous website ‘Yelp’, customers can give stars and write a review for a business. Also, Yelp shares its dataset. The dataset contains a lot of reviews with stars. Therefore, I will use this data to be my dataset. At first, I will randomly pick 10,000 reviews to be my training data. Another 2,000 to be validation data. Another 2,000 to be testing data. Secondly, I will parse all reviews to know how many words appear in these reviews. Thirdly, I will design a table where rows correspond to reviews and columns to numbers of words which exist in reviews. Fourthly, using the design of third step creates three tables, one made of training data, one made of validation data, and one made of testing data. Finally, using the table which contains training data and stars train a model. Also, use tables which contain testing data and validation data to test and validate the model.

All the labels which are stars of reviews are already a ground-truth, since these are real reviews from Yelp.