Lecture 7: Regression
Modeling Social Data, Spring 2019
Columbia University

Bhavya Shahi (bs3118)

March 8, 2019

1 Reproducibility

As discussed in the previous class, there is currently a crisis of reproducibility and replication. To counter this, there a few guidelines we should follow before we publish research.

- Read all the pertinent literature, as you may not be the first person to have this idea. Literature surveys also help in understanding our problem better
- Formulate your study
- Run a pilot to test the study
- Analyze the results of the study
- Revise your study based on this analysis
- Calculate the power of your study, to define the effect and population size needed
- Pre-regiser plans for your study online. This helps in reviewing your process, and ensures that you cannot lie about the results later on.
- Run your study
- Create a detailed and reproducible report using the makefile templates define during the pilot study
- Think critically about your results
- Disclose everything you did during the study, including data collection, preprocessing etc.

Other good practices for reproducibility include:

- Use Mechanical Turk to collect data since it is reliable and quick
- Running a pilot study first
- Creating a framework to analyse the data that can be easily reproduced based on the pilot study
- To ensure that the subjects in the study understand the questions, we should add some questions as sanity checks.
- We could also use timestamps of the data collected to remove any bots.

Also, we should remember:

Standard deviation: The variation in the data (this remains constant no matter how many samples we choose)

Standard error: This represents how far the sample mean is from the population mean. Thus, if we take more number of samples, we can reduce this.

2 Regression

Regression is used to estimate the relationship between variables. It can be used to summarize the data, predict future data, and explain relationships within the data. We need to find a balance between understanding out current data, and being able to generalize in a way that we can also predict future outcomes. The task is to find a function to do this. The notes below show the how the task is defined.

2.1 Task

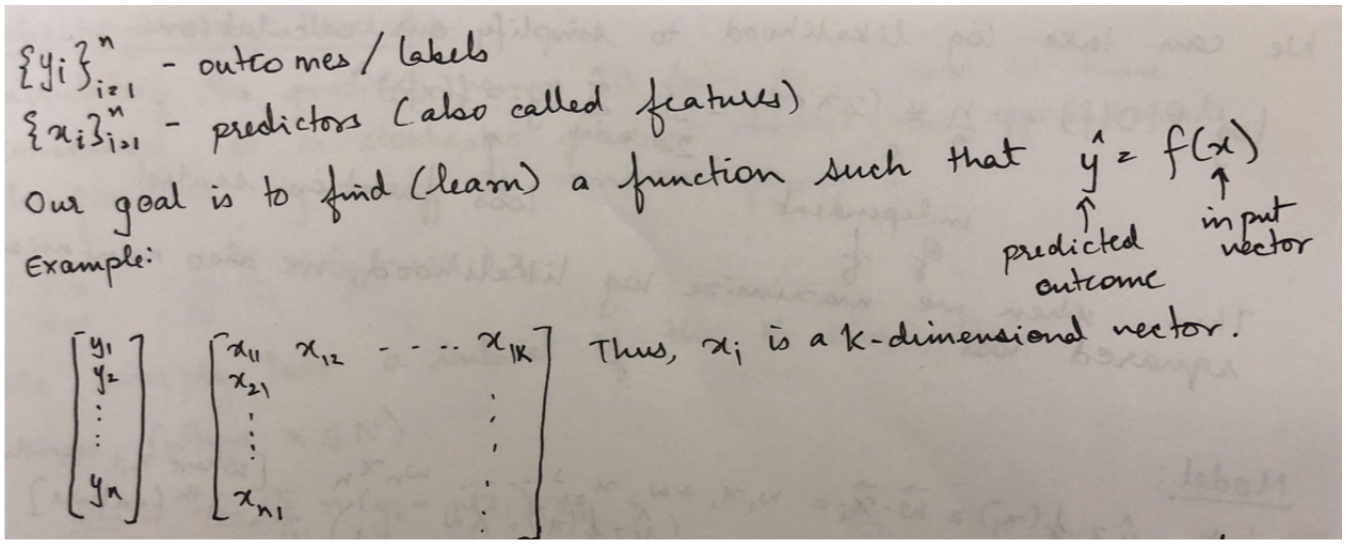

Figure 1: Defining the task

2.2 Loss functions

We then define loss functions to evaluate our function. The notes below show the how the task is defined.

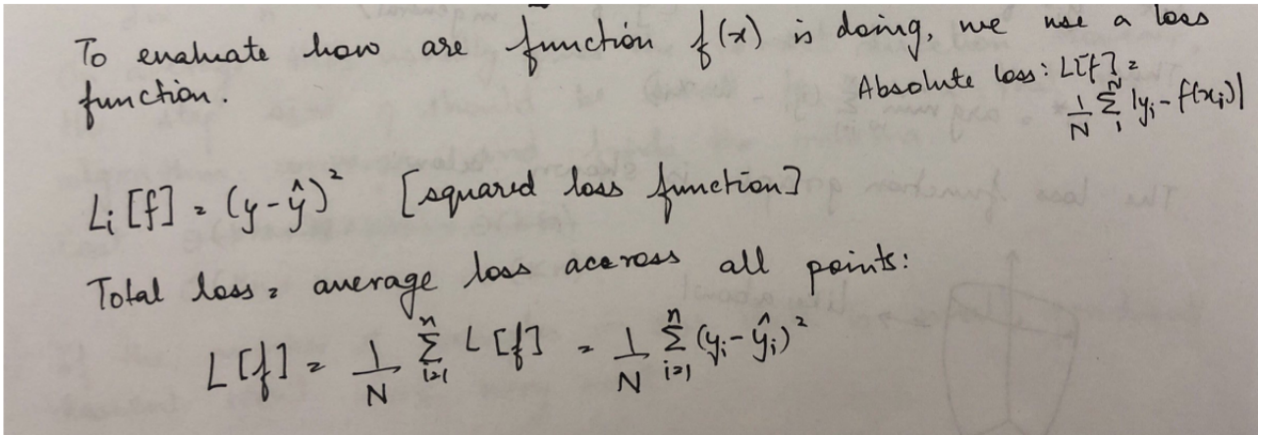

Figure 2: Define loss functions

2.3 MLE

Maximum Likelihood Estimator is used to find how to minimize loss.

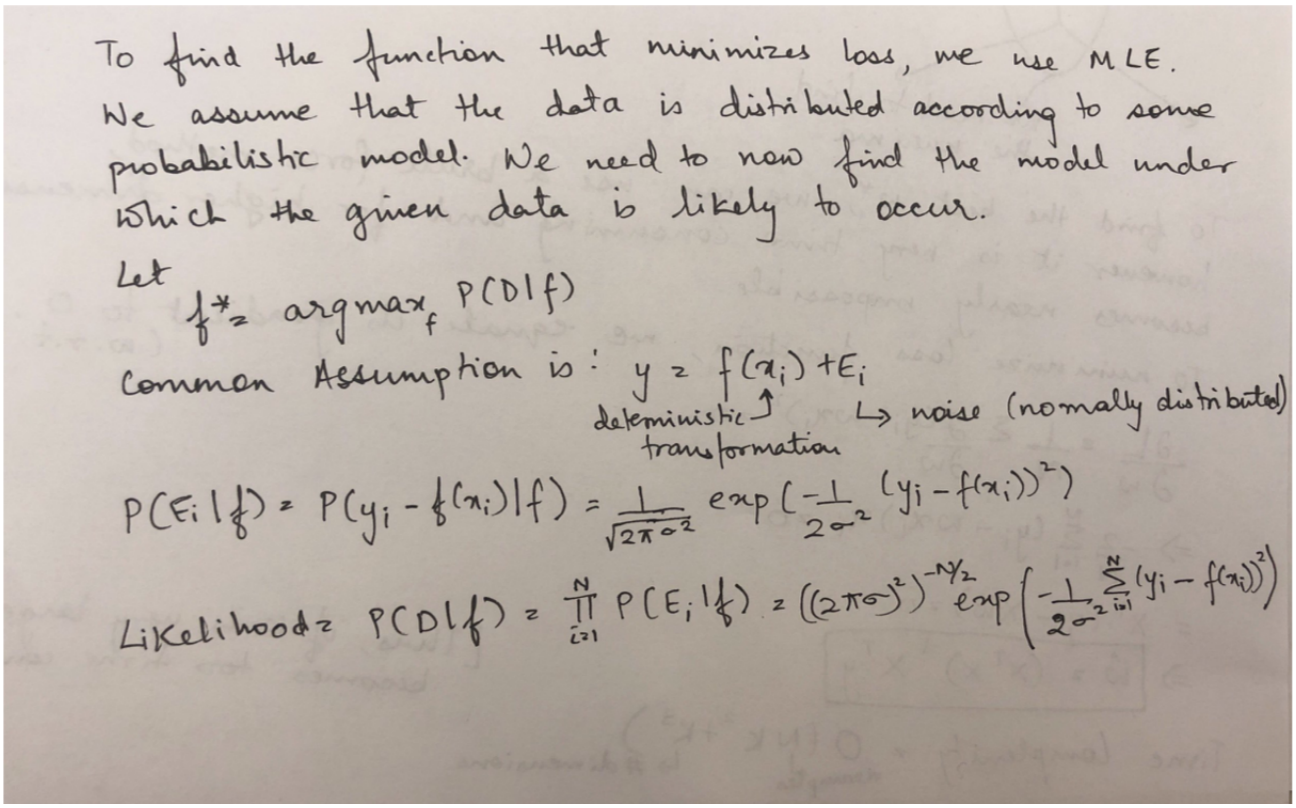

Figure 3: MLE

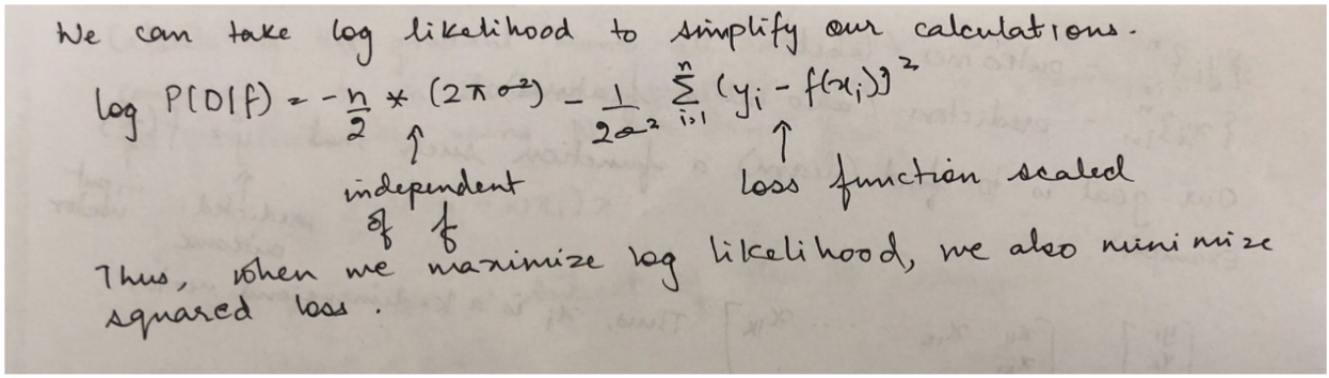

Figure 4: MLE

2.4 Finding f(x)

We then define the regression model in terms of weights. To find the best possible value of the weight vector, we can use three main methods: Normal equations (shown in Figure 5), Gradient Descent (Figure 6), and Stochastic Gradient Descent (shown in Figure 7)

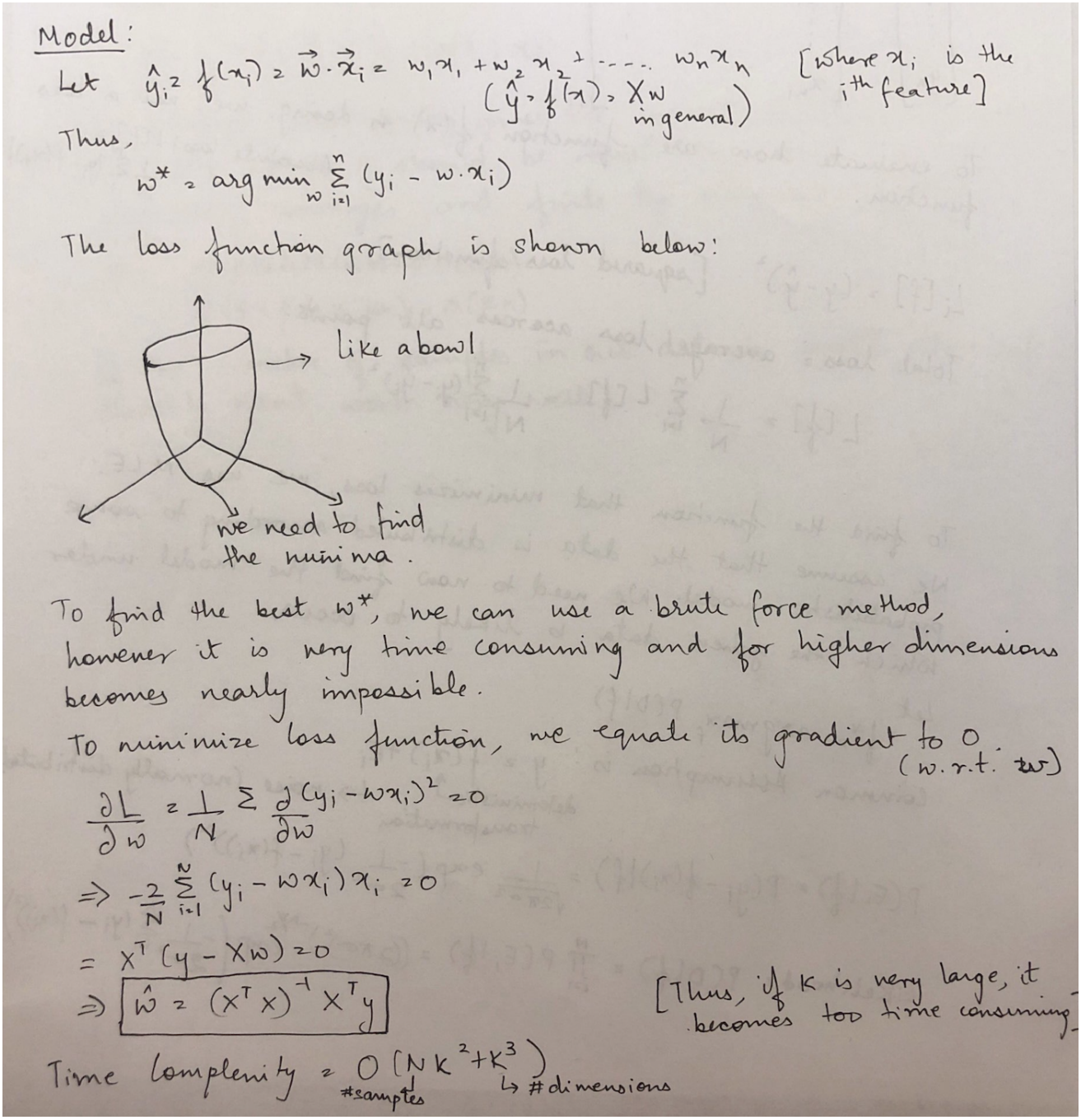

Figure 5: Defining and determining f(x)

2.5 Gradient Descent

Gradient descent can be used to estimate the w vector.

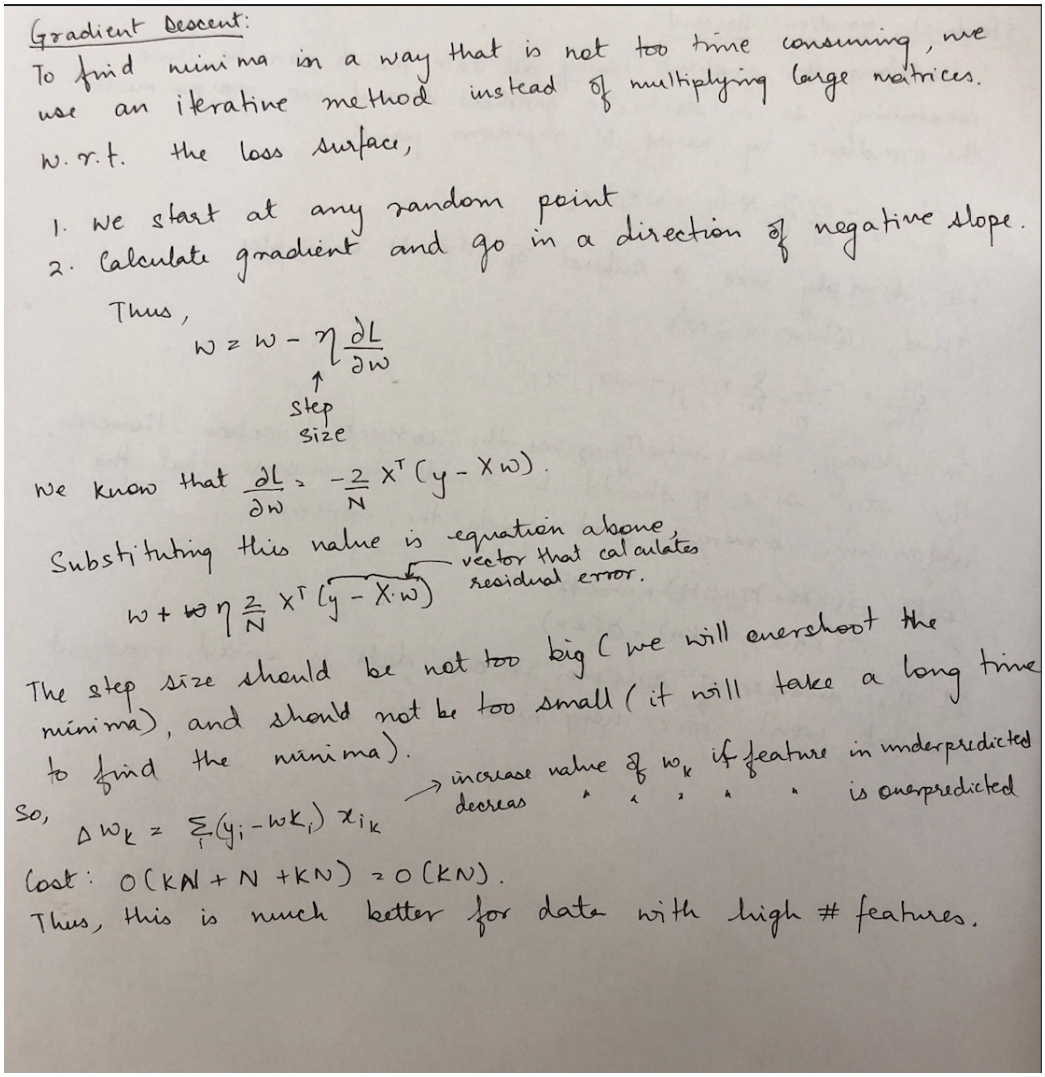

Figure 6: Gradient Descent

2.6 Stochastic gradient descent:

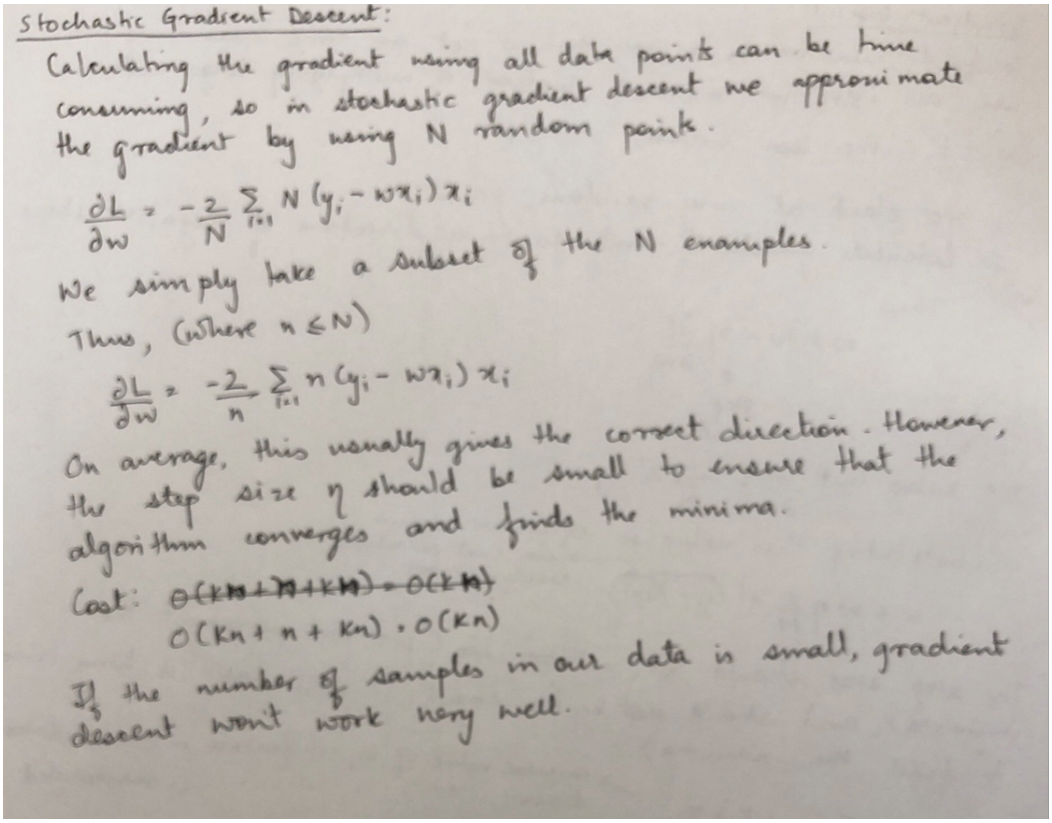

Figure 7: Stochastic Gradient Descent