

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 are 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-register 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 defined 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 our 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

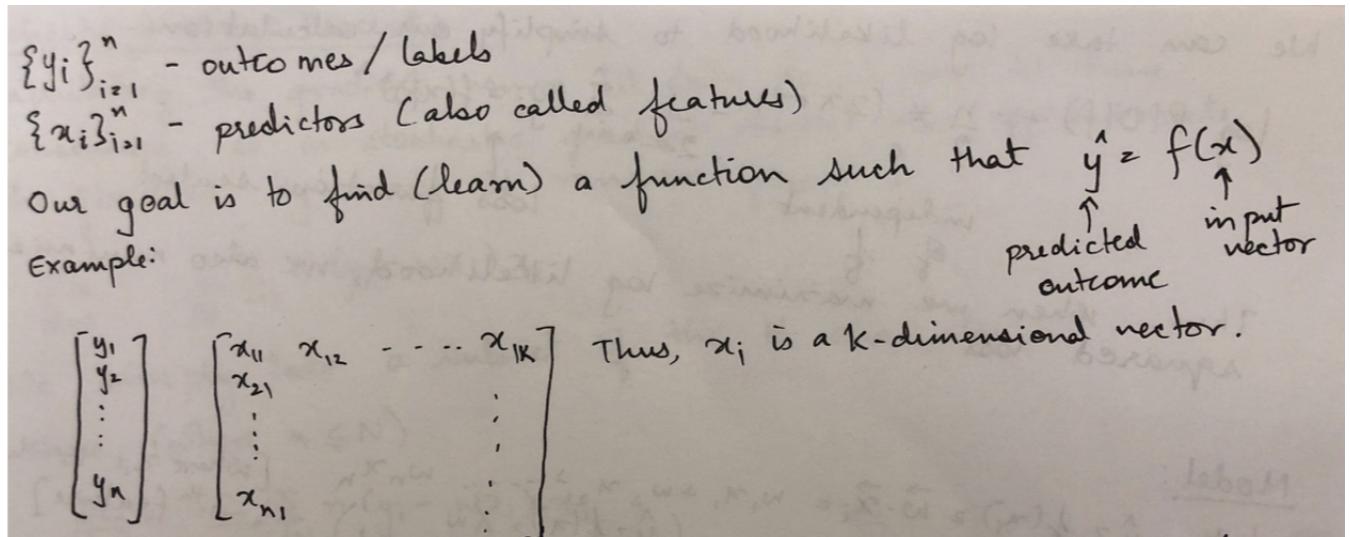


Figure 1: Defining the task

2.2 Loss functions

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

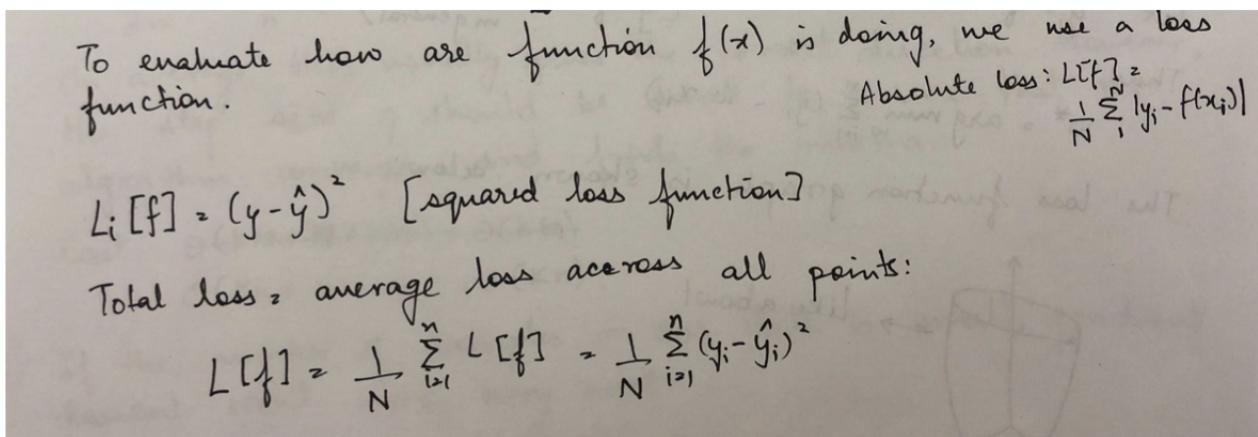


Figure 2: Define loss functions

2.3 MLE

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

To find the function that minimizes loss, we use MLE.

We assume that the data is distributed according to some probabilistic model. We need to now find the model under which the given data is likely to occur.

Let $f^* = \operatorname{argmax}_f P(D|f)$

Common Assumption is: $y = f(x_i) + E_i$

$\xrightarrow{\text{deterministic transformation}}$ $\xrightarrow{\text{noise (normally distributed)}}$

$$P(E_i | f) = P(y_i - f(x_i) | f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - f(x_i))^2\right)$$

$$\text{Likelihood} = P(D|f) = \prod_{i=1}^n P(E_i | f) = ((2\pi\sigma^2)^{-N})^{-N} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(x_i))^2\right)$$

Figure 3: MLE

We can take log likelihood to simplify our calculations.

$$\log P(D|f) = -\frac{n}{2} * (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(x_i))^2$$

\uparrow independent of f \uparrow loss function scaled

Thus, when we maximize log likelihood, we also minimize squared loss.

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)

Model:

Let $\hat{y}_i = f(x_i) = \vec{w} \cdot \vec{x}_i = w_1 x_1 + w_2 x_2 + \dots + w_n x_n$ [where x_i is the i^{th} feature]

($\hat{y} = f(x) = \vec{w} \cdot \vec{x}$ in general)

Thus,

$$w^* = \arg \min_w \sum_{i=1}^n (y_i - w \cdot x_i)$$

The loss function graph is shown below:

To find the best w^* , we can use a brute force method, however it is very time consuming and for higher dimensions becomes nearly impossible.

To minimize loss function, we equate its gradient to 0 (w.r.t. w)

$$\frac{\partial L}{\partial w} = \frac{1}{N} \sum_i \frac{\partial}{\partial w} (y_i - w \cdot x_i)^2 = 0$$

$$\Rightarrow -\frac{2}{N} \sum_{i=1}^N (y_i - w \cdot x_i) x_i = 0$$

$$= x^T (y - Xw) = 0$$

$$\Rightarrow \hat{w} = (x^T x)^{-1} x^T y$$

[Thus, if K is very large, it becomes too time consuming.]

Time complexity = $O(\frac{NK^2+K^3}{\text{#samples}})$ \hookrightarrow # dimensions

Figure 5: Defining and determining $f(x)$

2.5 Gradient Descent

Gradient descent can be used to estimate the w vector.

Gradient Descent:

To find minima in a way that is not too time consuming, we use an iterative method instead of multiplying large matrices. w.r.t. the loss surface,

1. We start at any random point
2. Calculate gradient and go in a direction of negative slope.

Thus,

$$w = w - \eta \frac{\partial L}{\partial w}$$

↑
Step
size

We know that $\frac{\partial L}{\partial w} = -\frac{2}{N} X^T (y - Xw)$.

Substituting this value in equation above, vector that calculates residual error.

$$w + w \eta \frac{2}{N} X^T (y - Xw)$$

The step size should be not too big (we will overshoot the minima), and should not be too small (it will take a long time to find the minima).

So, $\Delta w_k = \sum_i (y_i - w_k) x_{ik}$ → increase value of w_k if feature is underpredicted
decrease " " " " " " " if feature is overpredicted

Cost: $O(KN + N + KN) \approx O(KN)$.

Thus, this is much better for data with high # features.

Figure 6: Gradient Descent

2.6 Stochastic gradient descent:

Stochastic Gradient Descent:

Calculating the gradient using all data points can be time consuming, so in stochastic gradient descent we approximate the gradient by using N random points.

$$\frac{\partial L}{\partial w} = -\frac{2}{N} \sum_{i=1}^N (y_i - w x_i) x_i$$

We simply take a subset of the N examples.

Thus, (where $n \leq N$)

$$\frac{\partial L}{\partial w} = -\frac{2}{n} \sum_{i=1}^n (y_i - w x_i) x_i$$

On average, this usually gives the correct direction. However, the step size η should be small to ensure that the algorithm converges and finds the minima.

Cost: $O(Kn + n + Kn) = O(Kn)$
 $O(Kn + n + Kn) = O(Kn)$

If the number of samples in our data is small, gradient descent won't work very well.

Figure 7: Stochastic Gradient Descent