**WEEK 2 ANALYSIS AND OBSERVATIONS**

1. What are the modules that are imported?

->numpy(for mathematics and working with arrays): It provides fast and efficient operations on arrays of homogeneous data. Extends python into a high-level language for manipulating numerical data, similiar to MATLAB.

->matplotlib.pyplot(for plotting): is a collection of command style functions that make matplotlib work like MATLAB

->numpy.random: It uses a particular algorithm, called the Mersenne Twister, to generate pseudorandom numbers.

2. What are the models used?

a regression problem “which is predicting a continuous value like predicting prices of a house given the features of the house like size, number of rooms, etc”.

->Linear Regression: In [statistics](https://en.wikipedia.org/wiki/Statistics), linear regression is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modeling the relationship between a scalar response (or [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable)) and one or more [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or [independent variables](https://en.wikipedia.org/wiki/Independent_variable)). The case of one explanatory variable is called [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression).

->Polynomial Regression: In [statistics](https://en.wikipedia.org/wiki/Statistics), polynomial regression is a form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) in which the relationship between the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) x and the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) y is modelled as an nth degree [polynomial](https://en.wikipedia.org/wiki/Polynomial) in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of y

->logistic regression: In [statistics](https://en.wikipedia.org/wiki/Statistics), the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc... Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

3. What are the hyper-parameters of the models used?

In machine learning, a hyperparameter is a parameter whose value is set before the learning process begins.

->t0,t1:Here t0 and t1 values are set previously, being used in stochastic gradient descent

4. What are the Optimizers used? How to use them more efficiently.

optimizers shape and mold your model into its most accurate possible form by futzing with the weights. The loss function is the guide to the terrain, telling the optimizer when it’s moving in the right or wrong direction.

->gradient descent is the optimizer

->**Batch Gradient Descent or Vanilla Gradient Descent**

**->Stochastic Gradient Descent**

**->Mini batch Gradient Descent**

**->ways to use efficiently:**

* **The First Order Optimization** techniques are easy to compute and less time consuming , converging pretty fast on large data sets.
* **Second Order Techniques** are faster only when the **Second Order Derivative** is known otherwise, these methods are always slower and costly to compute in terms of both time and memory
* Optimizing the gradient descent using various algorithms such as: Momentum, Nesterov accelerated gradient, Adagrad

5. What is Stochastic Gradient Descent? How is it different from Gradient Descent?

->Stochastic gradient descent (often abbreviated SGD) is an [iterative method](https://en.m.wikipedia.org/wiki/Iterative_method) for [optimizing](https://en.m.wikipedia.org/wiki/Mathematical_optimization) an [objective function](https://en.m.wikipedia.org/wiki/Objective_function) with suitable smoothness properties (e.g. [differentiable](https://en.m.wikipedia.org/wiki/Differentiable_function) or [subdifferentiable](https://en.m.wikipedia.org/wiki/Subgradient_method)). It is called stochastic because the method uses randomly selected (or shuffled) samples to evaluate the gradients, hence SGD can be regarded as a [stochastic approximation](https://en.m.wikipedia.org/wiki/Stochastic_approximation) of [gradient descent](https://en.m.wikipedia.org/wiki/Gradient_descent) optimization.

->Unlike the batch gradient descent which computes the gradient using the whole dataset, because the SGD, also known as incremental gradient descent, tries to find minimums or maximums by iteration from a single randomly picked training example, the error is typically noisier than in gradient descent.

-> SGD often converges much faster compared to GD but the error function is not as well minimized as in the case of GD. Often in most cases, the close approximation that you get in SGD for the parameter values are enough because they reach the optimal values and keep oscillating there.

6. What is Learning Schedule?

-> Learning rate is a hyper-parameter that controls how much we are adjusting the weights of our network with respect the loss gradient. The lower the value, the slower we travel along the downward slope. Typically learning rates are configured naively at random by the user.

7. What is Mini-Batch Gradient Descent? What are its benefits over regular Gradient Descent?

-> Mini-batch gradient descent is a variation of the gradient descent algorithm that splits the training dataset into small batches that are used to calculate model error and update model coefficients. Implementations may choose to sum the gradient over the mini-batch which further reduces the variance of the gradient.

8. Benefits and Drawbacks of Polynomial Regression over Linear Regression.

->Advantages: Works on any size of data. Works best for non-linear problems.

->Disadvantages: Need to be careful about choosing the right polynomial degree variance/bias tradeoff.

9. Benefits of Standard Scalar.

-> The idea behind Standard Scaler is that it will transform your data, such that the distribution will have a mean value of 0 and a standard deviation of 1.

->First step of data pre-processing

->helps to normalize the data within a particular range.

->Also helps in speeding up the calculations in an algorithm.

10. Difference between Logistic Regression and Linear Regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BASIS FOR COMPARISON | LINEAR REGRESSION | LOGISTIC REGRESSION   |  | | --- | |  | |
| Basic | The data is modelled using a straight line. | The probability of some obtained event is represented as a linear function of a combination of predictor variables. |
| Linear relationship between dependent and independent variables | Is required | Not required |
| The independent variable | Could be correlated with each other. (Specially in multiple linear regression) | Should not be correlated with each other (no multicollinearity exist). |
| Target | Target is an interval variable. | Target is a discrete(binary or ordinal) variable. |

11. What is Softmax?

-> The Softmax regression is a form of logistic regression that normalizes an input value into a vector of values that follows a probability distribution whose total sums up to 1.

-> Softmax is implemented through a neural network layer just before the output layer. The Softmax layer must have the same number of nodes as the output layer

12. Any additional insights and Analysis you want to perform on your own.

-> I would like to understand in depth the strength of impact of multiple independent variables on a dependent variable.

->More depth into softmax and analysis the proper implementation of softmax regression and its working