1. What do you understand by binomial logistic regression and multinomial logistic regression?
2. What are the assumptions made of logistic regression on data?
3. Why linear regression cannot be used to classify data?
4. What is the significance of sigmoid function in logistic regression?
5. What is the decision boundary?
6. How does the sigmoid function help in mitigating the problems associated due to outliers?
7. Can the cost function used in linear regression work in logistic regression?
8. What is squashing in the context of logistic regression?
9. Scaling is required in logistic regression?If yes then why?
10. Explain AUC and ROC curve.?

Solutions

1)Binomial logistic regression is type of classification technique which classifies 2 classes.Multinomial logistic regression handles more than 2 classes classification task.

2)Following are the assumptions made by logistic regression on the data

1. The data is linearly separable.
2. The 2 classes can be separated by straight line.
3. Either one of the class is assumed to be positive class and represented by +1 and vice versa for other class.
4. The points lying above the plane or straight line will have positive distance.
5. The points lying below the plane or straight line will have negative distance.

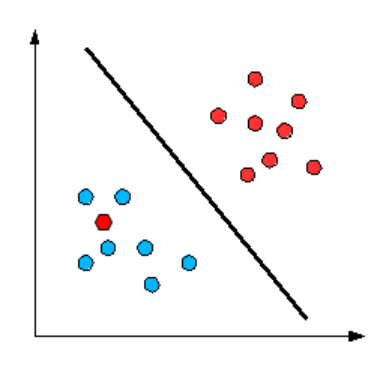
3)Linear regression tries to draw straight line in the given data and in case of classification task ,it tries to separate the 2 classes.This line is called as best fit line and this line is super sensitive to outliers i.e whenever we have outliers in the data,the bestfit line keeps on changing making wrong prediction i.e more misclassification.We can use Linear regression to classify points but we will end up getting more misclassified points.

4)The significance of sigmoid equation as below

Sigmoid function is used to eliminate the effect of outliers. The reasons for choosing the sigmoid functions are as follows:

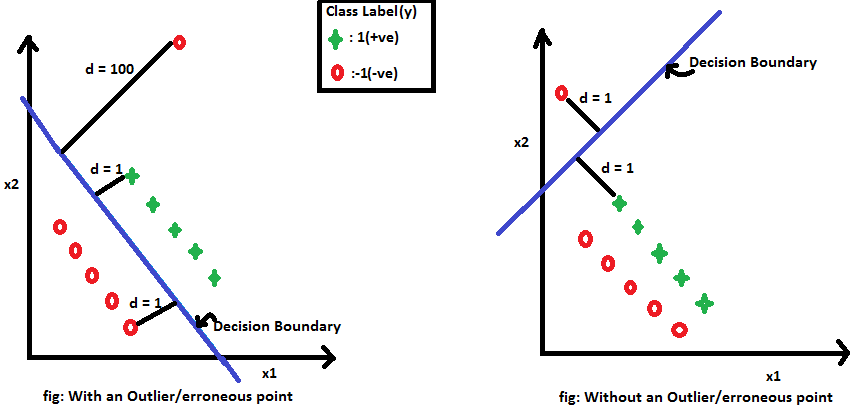
1. Probabilistic Inference: If z=y\*w^T\*xi=0, it means d=w^T\*xi is 0, i.e, the shortest distance of the point from the plane is zero. Now, if d is 0 it means the point lies on the hyperplane itself. Again, if z=0 then P(z)=0.5.
2. Function is differentiable at every point: An optimization algorithm like Gradient Descent is used to find the values of w and b. As sigmoid function is differentiable so the minima and maxima can be easily calculated.

5) A line or a hyperplane that separates the classes is called a decision boundary. The goal of logistic regression, as with any classifier, is to figure out some way to split the data to allow for an accurate prediction of a given observation’s class using the information present in the features. (For instance, if we were examining the Iris flower dataset, our classifier would figure out some method to split the data based on the following: sepal length, sepal width, petal length, petal width.) In the case of a generic two-dimensional example, the split might look something like this.



6) The sigmoid function plays an important role in mitigating the problem of outliers. Let’s take a very simple example where we will see how sum of signed distances (yi\*w^t\*xi) can be impacted by an erroneous/outlier points and we need to come up with another formulation which is less impacted by outlier.

Suppose in the figure left, the distance (d) from any point to decision boundary is 1 for all -ve side of decision boundary points and +ve side of decision boundary points, except an outlier point which is in the +ve side of the decision boundary and the distance is 100. If we compute the signed distance then it will be -90. In the figure right, the distance (d) from any point to decision boundary is 1 and their distances from each other is also 1. If we compute the signed distance then it will be 1. So, we have 5 miss-classified points (point is -ve but are in +ve side of the decision boundary) and sum of signed distance is -90. In figure left, we have 1 miss-classified point and sum of signed distance is 1. And we want to maximize the sum of signed distances which is 1 in this case. So, If we choose sum of signed distance, in the presence of outlier, our prediction may not be correct and we end up with worst model.



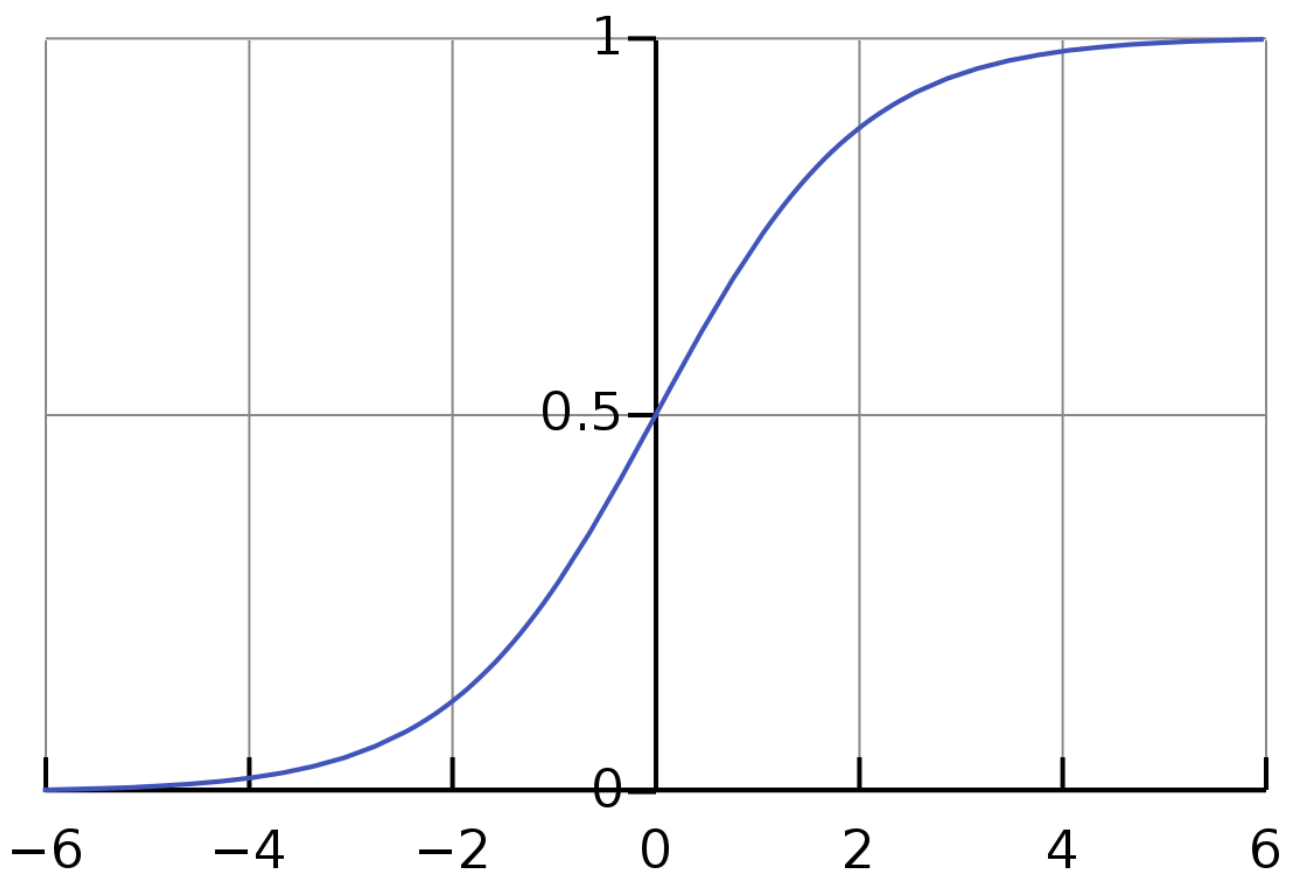
So, to avoid this problem we need another function that can be more robust than the maximizing signed distances . Such function we use here is called the sigmoid function and is define as



So, we need to maximize the sigmoid function which is defined as

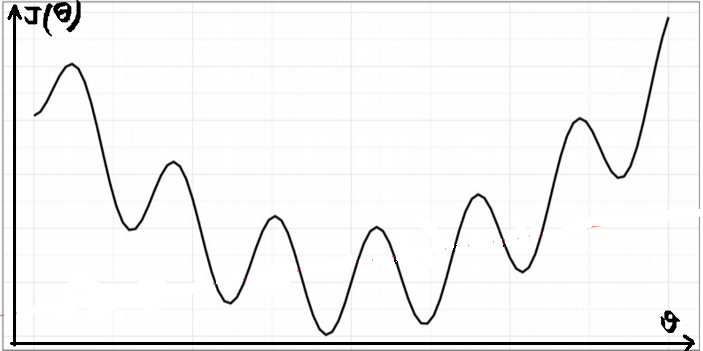


Thus we take the output(z) of the linear equation and give to the function g(x) which returns a squashed value h, the value h will lie in the range of 0 to 1. To understand how sigmoid function squashes the values within the range, let’s visualize the graph of the sigmoid function.



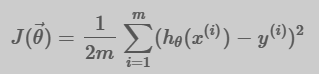
As we can see from the graph, the sigmoid function becomes asymptote to y=1 for positive values of x and becomes asymptote to y=0 for negative values of x.

7)The cost function J(θ) used in linear regression cannot work with logistic regression. In linear regression, we used the squared error mechanism. Unfortunately for logistic regression, such a cost function produces a nonconvex space that is not ideal for optimization. There will exist many local optima on which our optimization algorithm might prematurely converge before finding the true minimum.

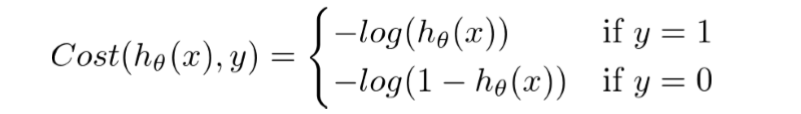


This strange outcome is due to the fact that in logistic regression we have the sigmoid function , which is non-linear (i.e. not a line). As a result the gradient descent algorithm might get stuck in a local minimum point. Using the Maximum Likelihood Estimator from statistics, we can obtain the following cost function which produces a convex space friendly for optimization. This function is known as the **binary cross-entropy loss**. These cost functions return high costs for incorrect predictions.

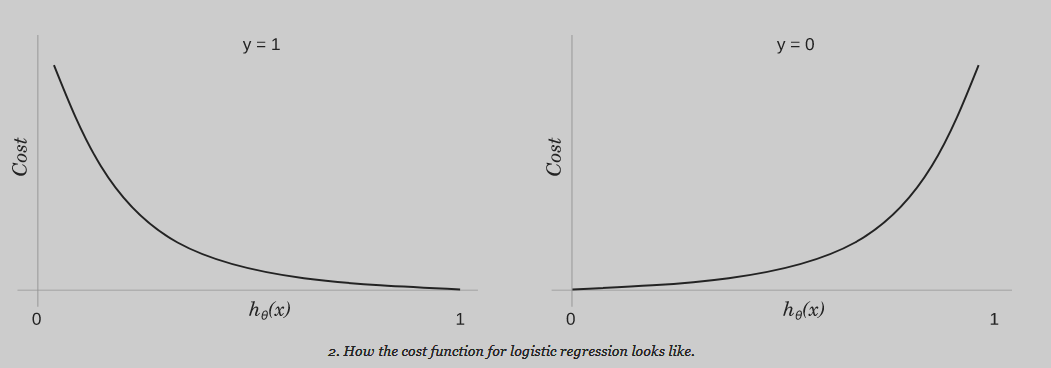
The cost function we used in linear regression was:



For logistic regression, the Cost function is defined as:



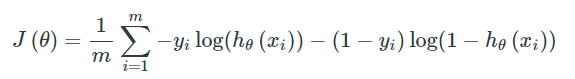
In case y=1, the output (i.e. the cost to pay) approaches to 0 as hθ(x) approaches to 1. Conversely, the cost to pay grows to infinity as hθ(x) approaches to 0. You can clearly see it in the plot below, left side. This is a desirable property: we want a bigger penalty as the algorithm predicts something far away from the actual value. If the label is y=1 but the algorithm predicts hθ(x)=0, the outcome is completely wrong.

Conversely, the same intuition applies when y=0, depicted in the plot below, right side. Bigger penalties when the label is y=0 but the algorithm predicts hθ(x)=1.

In brief the cost function can be written as:



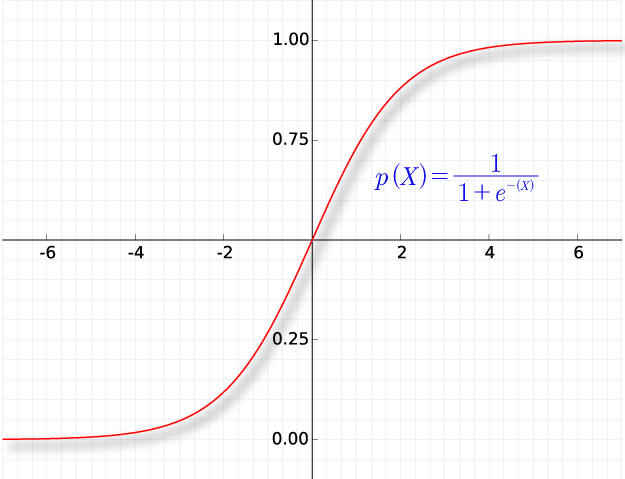
since the second term will be zero when y=1 and the first term will be zero when y=0. Substituting this cost into our overall cost function we obtain:



8)Squashing function maps the whole real axis into finite interval.

As Zi goes from -∞ to +∞, f(Zi) goes from A to B.

Sigmoid Function :



We will be using sigmoid function to squash the value between 0 and 1.

Usually, the predictions in the classification problem are probability values. So we don’t want our model to predict the probability value below 0 or above 1. Sigmoid function helps to achieve that.

9)Yes scaling is must while apply logistic regression to any data.As the cost function of logistic regression involves distance calculation hence scaling becomes important.

10)ROC:-Receiver Operating The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values.

\* TPR:-True Positive Rate/Recall TP/TP+FN.Sensitivity tells us what proportion of the positive class got correctly classified.

\* A simple example would be to determine what proportion of the actual sick people were correctly detected by the model.

\* False Positive Rate:-FP/FP+TN.FPR tells us what proportion of the negative class got incorrectly classified by the classifier.

\* ROC is used to select threshold values.

\* This concept is only applicable on Binary classification.By default threshold value is 0.5.

How Does the AUC-ROC Curve Work?

In a ROC curve, a higher X-axis value indicates a higher number of False positives than True negatives. While a higher Y-axis value indicates a higher number of True positives than False negatives. So, the choice of the threshold depends on the ability to balance between False positives and False negatives.

