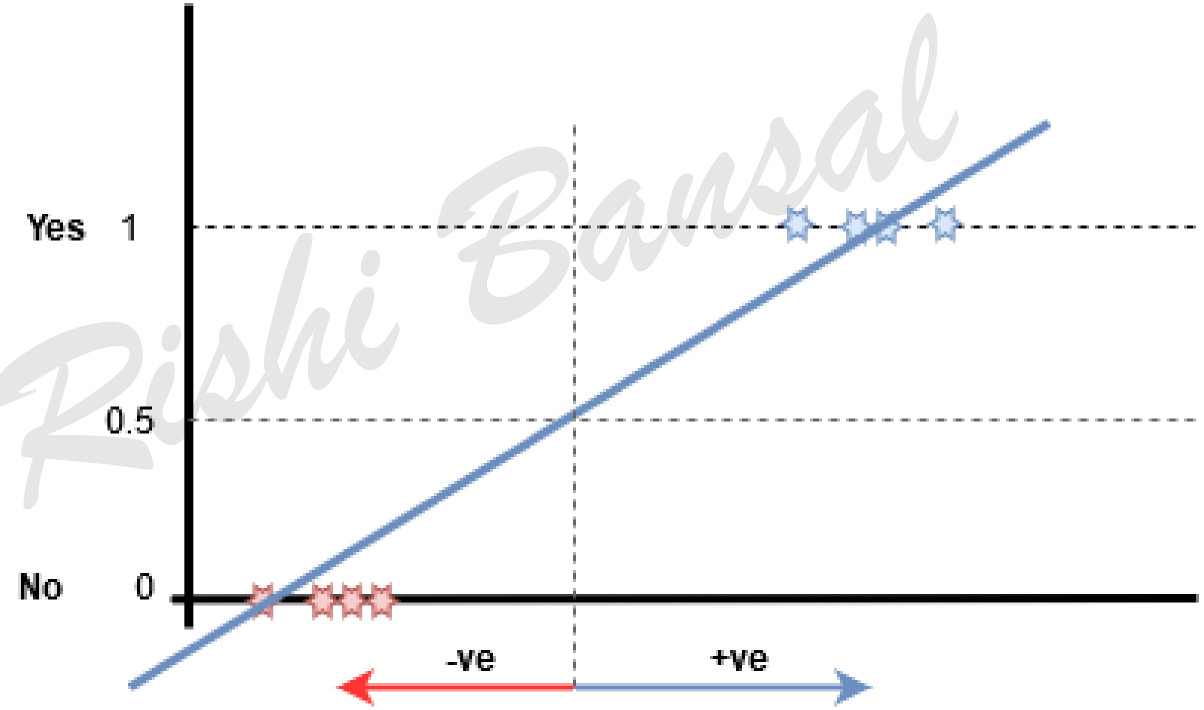
Classification

Classification

- A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”

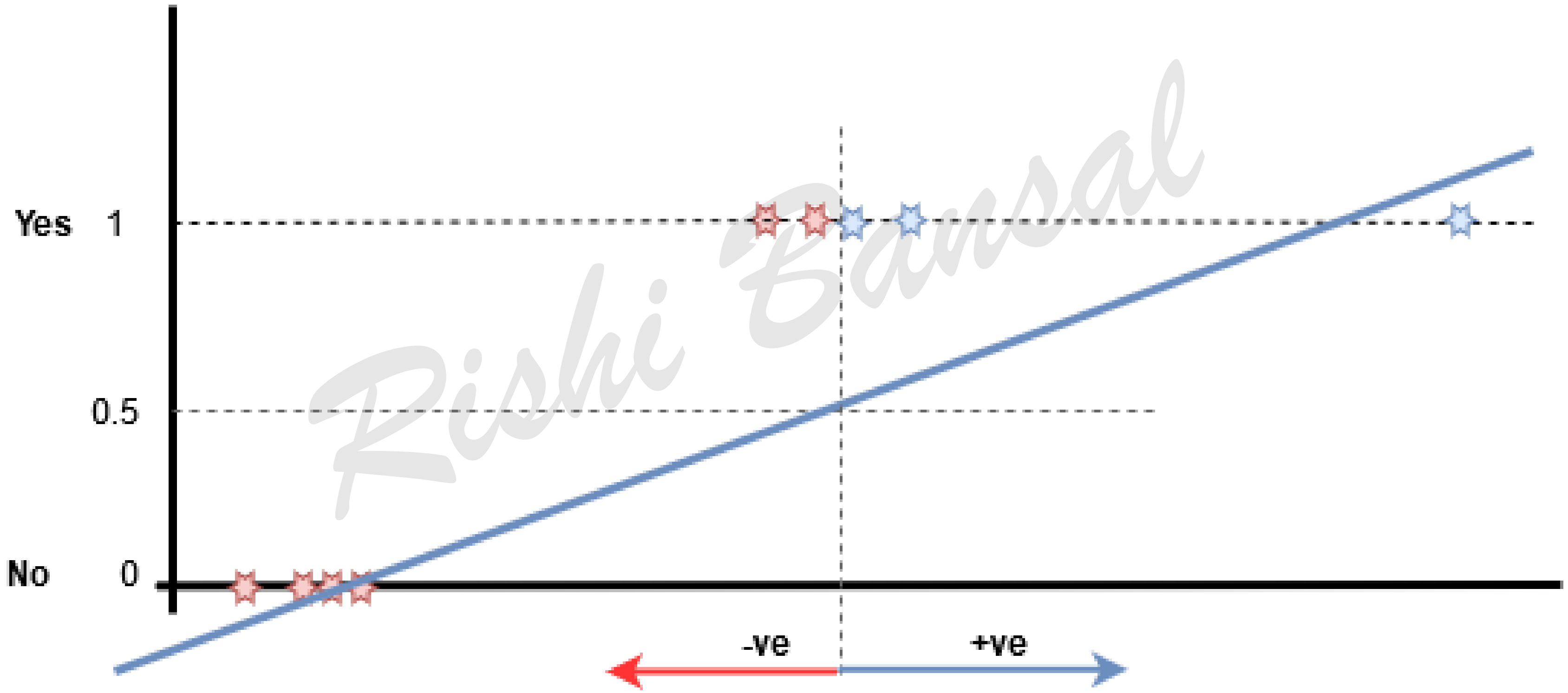
Issue with Linear Regression

Def: Suppose we need to decide whether on tumor size vs its malignancy. Since it is a classification problem, on plotting, all the values will lie on 0 and 1. And if we fit best found regression line, by setting the threshold at 0.5, it will give a descent result.



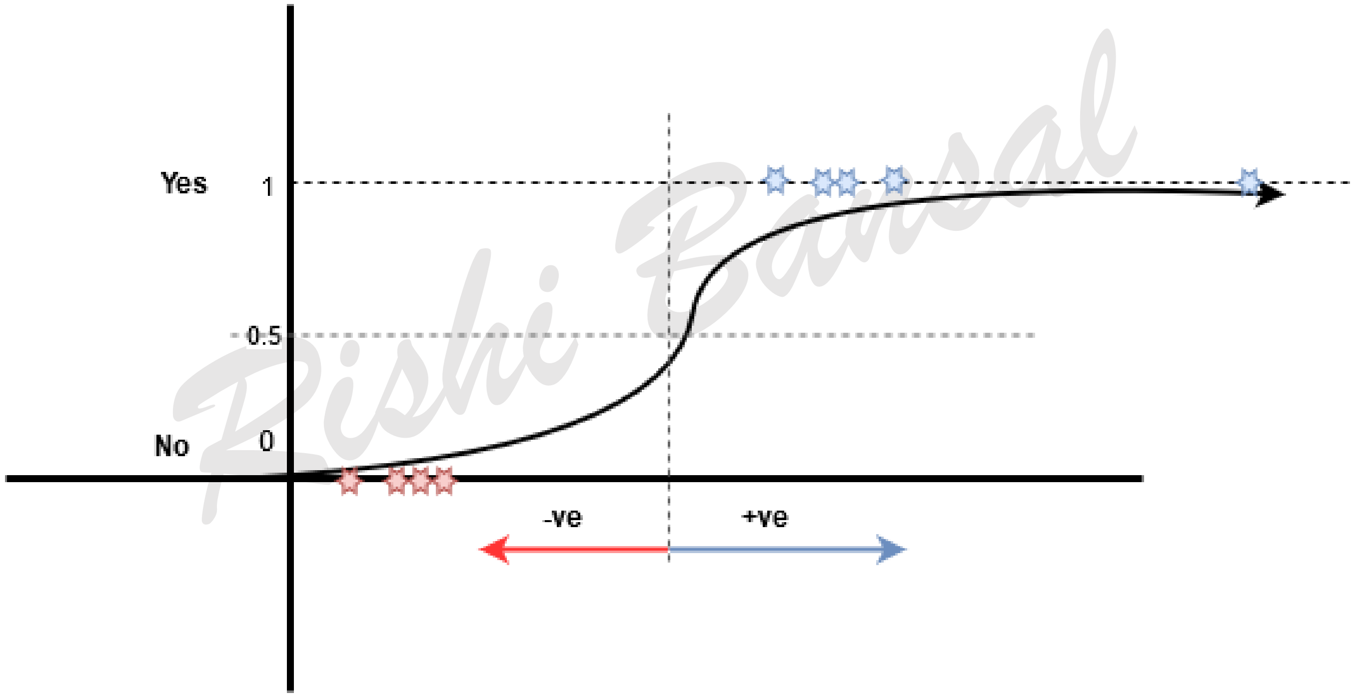
Issue with Linear Regression

- But if we have an outlier, it will go horribly wrong
- Because of one outlier, whole linear regression prediction is going wrong



Logistic Regression

- Logistic function is a Sigmoid function, which takes real value between zero and one.
- If we plot sigmoid function, the graph will be S curve. When there is an outlier, sigmoid function takes care of it.



Logistic Regression

- Linear regression assumes that the data follows a linear function
- Logistic regression models the data using the sigmoid function

- Linear R: y = bo + b1x
- Sigmoid Func: input vector x has n features and we have a corresponding parameter θj for each feature xj we can rewrite the logistic function as:

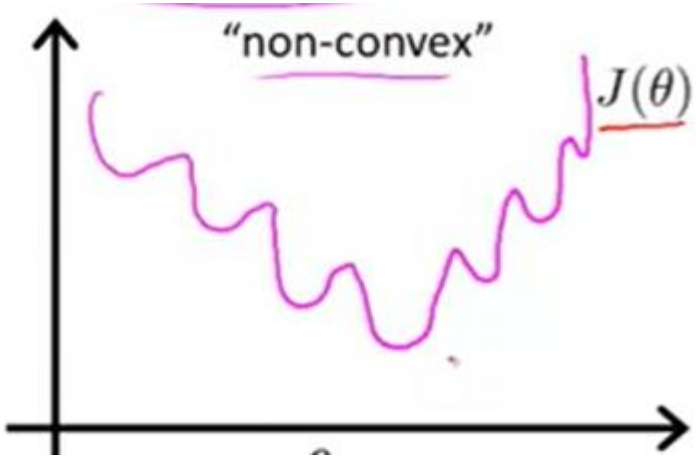
$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Cost Function

- Linear Regression

Cost function

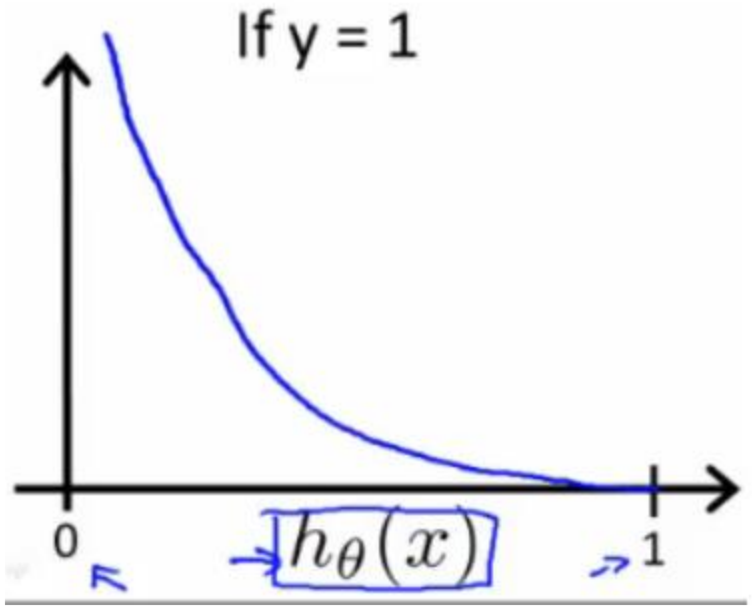
- Linear regression: $J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

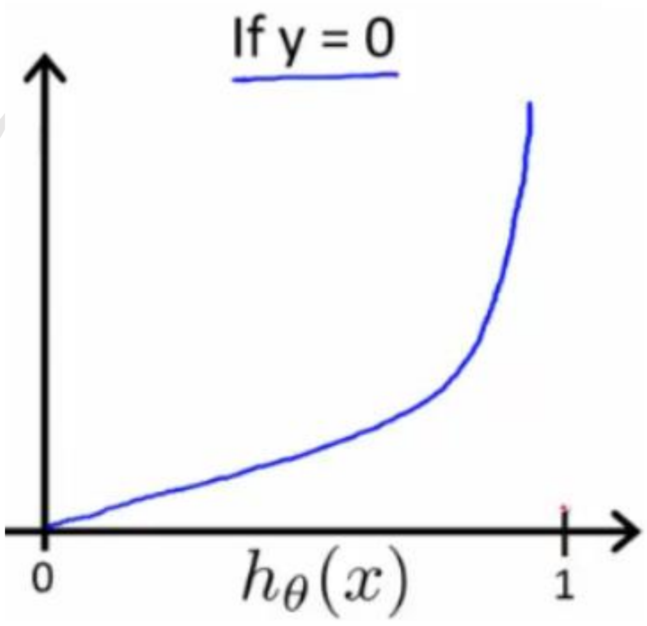


Cost Function

- Cost(hθ(x),y)= {−log(hθ(x))   if y = 1
  −log(1−hθ(x))   if y = 0

Cost(hθ(x),y)=−ylog(hθ(x)) − (1−y)log(1−hθ(x))





Cost Function

- Cost Function:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

cost($h_{\theta,}$ (x),y) = -ylog(h_θ(x)) - (1-y)log(1- h_θ(x))

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Cross Validation

- Here we are influenced by Test Data
- Evaluate Training without Test Data

- Use Test data only for final test
- **Issue**:
- How do we choose the validation set?
- Enough data?
- Mitigate overfitting?

K – Fold Cross Validation

Steps:

- For each fold, determine best hyper parameter value
- Now take average best value of HP as model HP value
- This will provide HP value good for all set of data

Hyper-Parameters

- One which is set manually before learning process begins.
- Hyper-parameters are data dependent & many times need experiments to find the best
- GridSearch is used to find the best hyper-parameters
- hyperparameter is a parameter whose value is used to control the learning process

Examples:

- The learning rate for training a neural network.
- The C and sigma hyperparameters for support vector machines.
- The k in k-nearest neighbors.

Grid Search

- Grid-search is used to find the optimal hyperparameters of a model which results in the most 'accurate' predictions