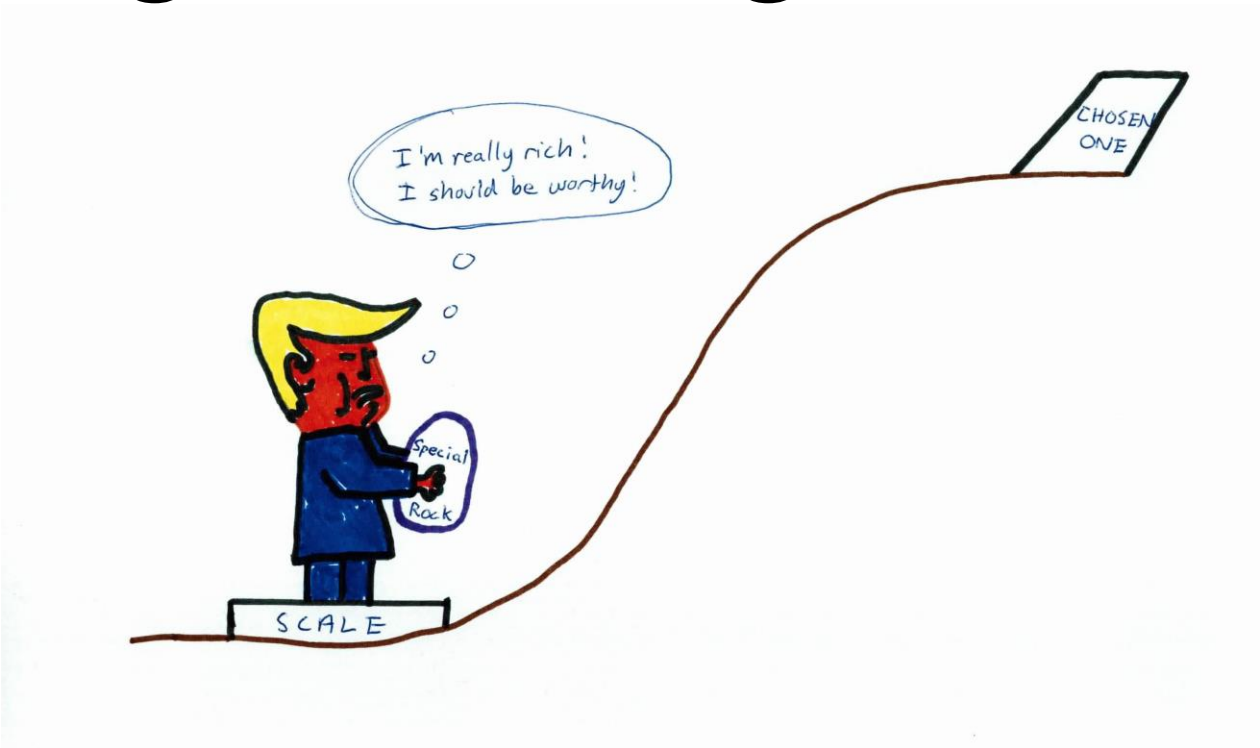
# Logistic Regression



# Categorical Response Variables

Examples

Whether or not a person smokes

Binary Response

$$Y = \begin{cases} \text{Non} - \text{smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment

$$Y = \begin{cases} \text{Survives} \\ \text{Dies} \end{cases}$$

Opinion poll responses

Ordinal Response

$$Y = \begin{cases} \text{Agree} \\ \text{Neutral} \\ \text{Disagree} \end{cases}$$

# Introduction

- Logistic regression is a statistical and machine learning technique for classifying records of a dataset, based on the values of the input fields.

- Let's say we have a telecommunication dataset which you'll use to build a model based on logistic regression for predicting customer churn, using **the given features.**

Dependent Variable

INDEPENDENT VARIABLES

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

# Logistic Regression Applications

- **Predict the probability of a person having a heart attack** within a specified time period, based on our knowledge of the person's age, sex, and body mass index.
- **Predict the chance of mortality** in an injured patient
- **Predict whether a patient has a given disease**, such as diabetes, based on observed characteristics of that patient
- **Predict the likelihood of a customer purchasing a product** or halting a subscription.
- **Predict the probability of failure of a given process, system, or product**.

Note:- All of these applications, we not only predict the class of each case, **we also measure the probability of a case belonging to a specific class.**

# Logistic Regression Model

- Ideally, a logistic regression model, so called ŷ, can **predict that the class of a customer is 1, given its features x (probability of customer falling in a particular class).**

$$\hat{y} = P(y = 1|x)$$

For class of customer =0

$$\hat{y} = P(y = 0|x) = 1 - P(y = 1|x)$$

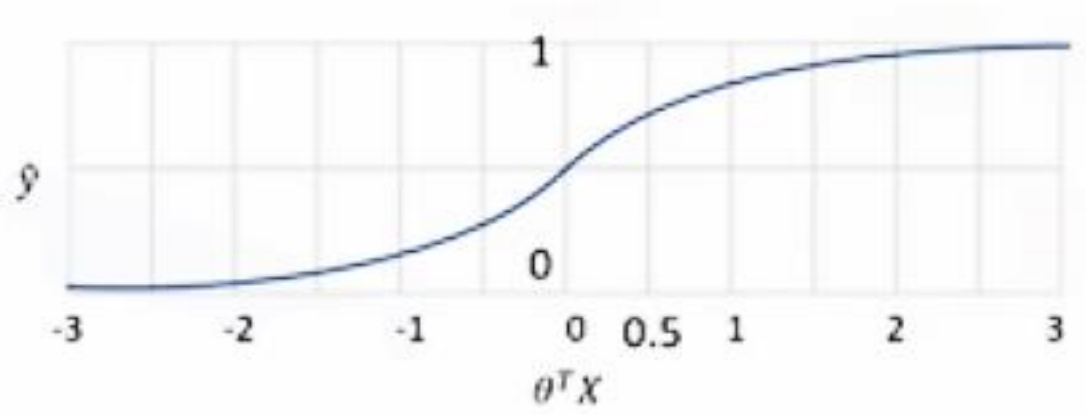
x (independent variables) → y (dependent variable)

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

# Logistic Regression Model

- To build such model, Instead of using $\theta^T X$ we use a specific function called sigmoid.

- $\sigma(\theta^T X)$ gives us the probability of a point belonging to a class, instead of the value of y directly.

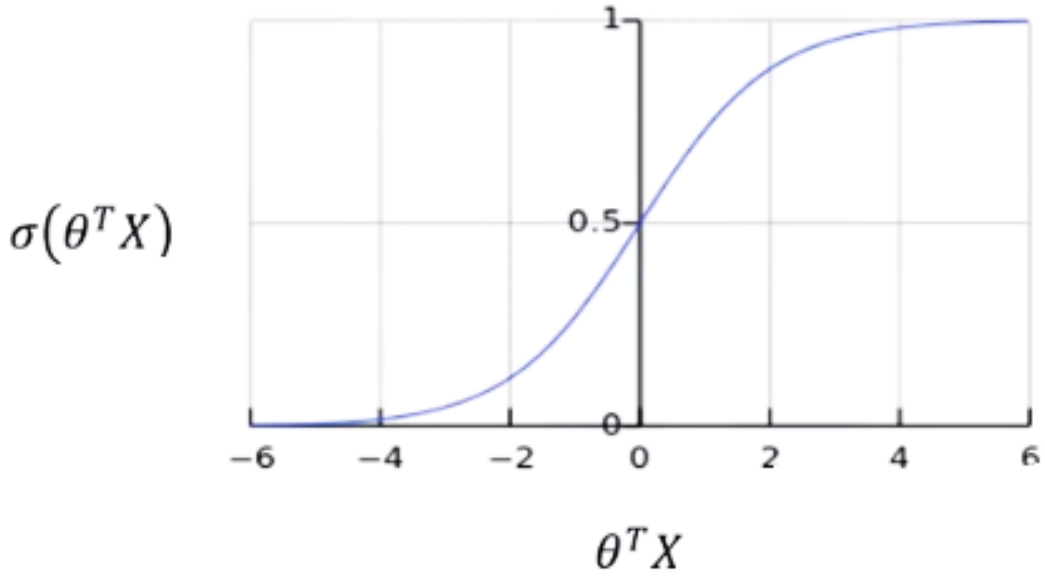



$$\hat{y} = \sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \cdots)$$

# Logistic Function

- **Sigmoid function, ( the logistic function),** resembles the step function and is used by the following expression in the logistic regression.

$$\sigma(\theta^T X) = \frac{1}{1+e^{-\theta^T X}}$$

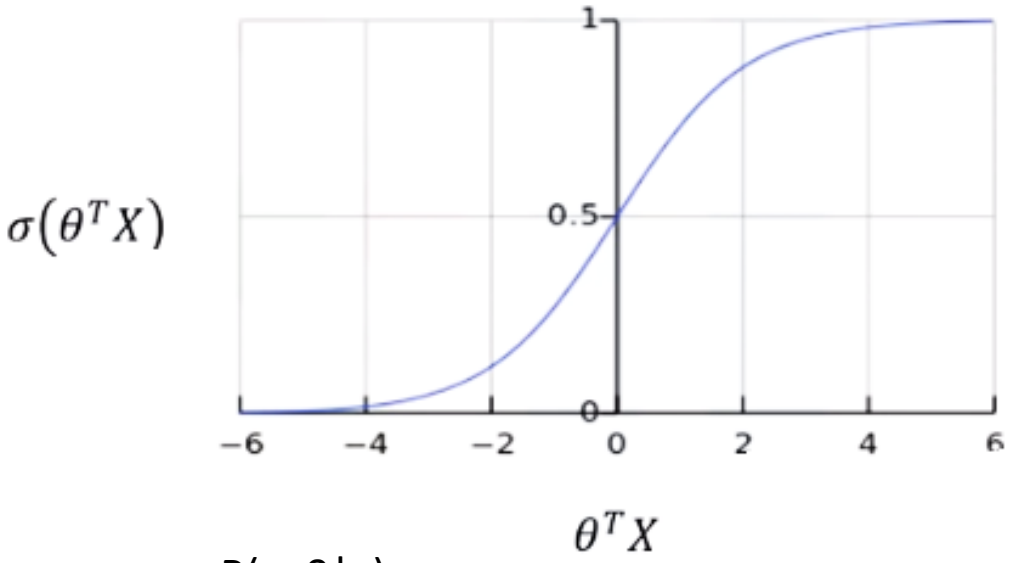


# Sigmoid Function

- When $\theta^T X$ goes bigger, sigmoid function gets closer to 1, the P(y=1|x), goes up and when $\theta^T X$ goes very small, sigmoid function gets closer to 0, thus, the P(y=1|x), goes down.

- Sigmoid function's output is always between 0 and 1, which makes it proper to interpret the results as probabilities.

P(y=1|x)

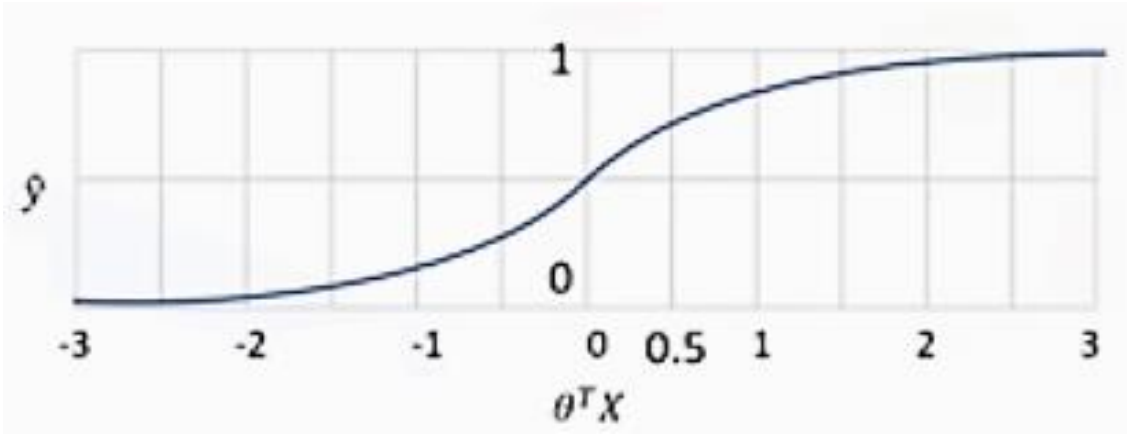



P(y=0|x)

# Output of Model with Sigmoid function

- Ex: the probability of a customer staying with the company can be shown as probability of churn equals 1 given a customer's income and age, assume, 0.7.

  P(churn=1|income,age) = 0.7




- And the probability of churn is 0, for the same customer:

  P(churn=0|income,age) = 1-0.7=0.3

$$\hat{y} = \sigma(\theta^T X) = P(y = 1|x)$$

# Model Build?

- **How can we build such model ?**

- First step towards building such model is to find $\theta$, which can find through the training process.

# Training Process

- **Step 1:** Initialize $\theta$ vector with random values, assume [ −1, 2].

- **Step 2:** Calculate the model output, which is σ($\theta^T X$), for a sample customer.
  - X in $\theta^T X$ is the feature vector values. Ex: the age and income of the customer, assume, [2, 5].
  - **θ is the confidence or weight** that you've set in the previous step.
  - The probability that the customer belongs to class 1 is:

$$\hat{y} = \sigma(\theta^T X) = \sigma([-1, 2] * [2, 5]) = 0.7$$

- **Step 3:** Compare the output of our model, ŷ, with the actual label of the customer, assume, 1 for churn. Ex: Model's error = 1-0.7 = 0.3

  - This is the error for only one customer out of all the customers in the training set.

# Training Process

**Step 4:** Calculate the error for all customers.

- Add up to find total error, which is the cost of your model,
- Cost function (error of the model) is the difference between the actual and the model's predicted values.

Therefore, the lower the cost, the better the model is at estimating the customer's labels correctly.

**We must try to minimize this cost !**

# Training Process

- **Step 5:** But, because the initial values for θ were chosen randomly, it's very likely that the cost function is very high. So, we change the $\theta$ in such a way to hopefully reduce the total cost.

- **Step 6:** After changing the values of θ, we go back to step 2 to start another iteration and calculate the cost of the model again **until the cost is low enough**.

# Estimation of θ

- There are different ways to change the values of θ, but one of the most popular ways is **gradient descent.**

- We must continue iterations and training of our model till we reach desired level of accuracy and stop it when it’s satisfactory.

# Loss Function

- Usually, the square of this equation is used because of the possibility of the negative result, and for the sake of simplicity, half of this value is considered as the cost function, through the derivative process.

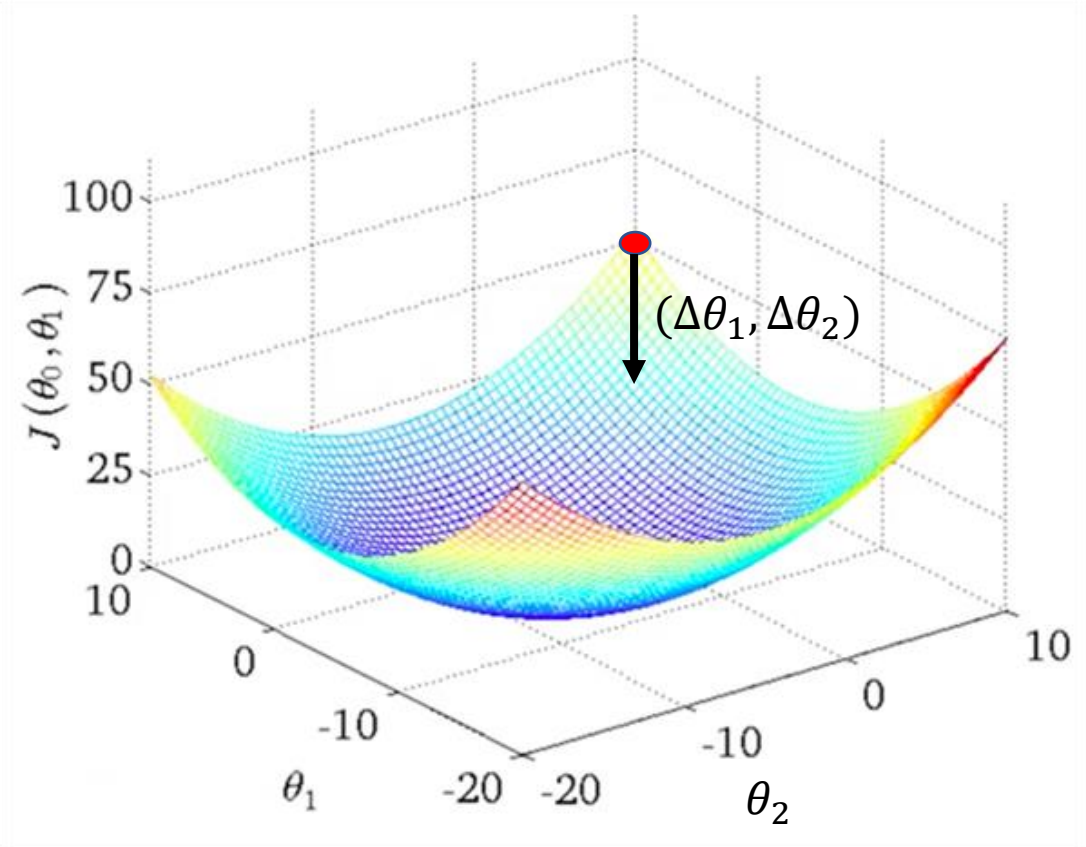
$$Cost(\hat{y}, y) = \frac{1}{2}(\sigma(\theta^T X) - y)^2$$

- Now, we can write the cost function for all the samples in our training set; for example, **for all customers, we can write it as the average sum of the cost functions of all cases.**

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

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} y^i \log\left(\hat{y}^i\right) + \left(1 - y^i\right)\log(1 - \hat{y}^i)$$

# Gradient Descent




- if we plot the cost function based on all possible values of $\theta 1$, $\theta 2$,
- **We call it "error curve" or "error bowl" of cost function.**
- It represents the error value for different values of parameters

# Hands On




# Thank You