LOGISTIC REGRESSION

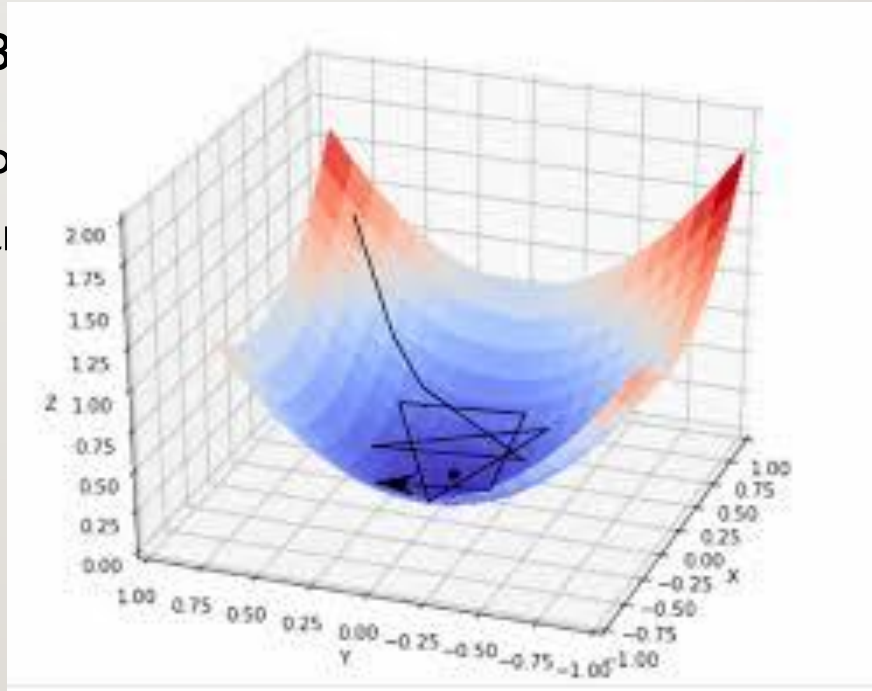
GRADIENT DESCENT

- This is also called unconstrained optimization
- Almost all ML algorithms involve solving an optimization problem.
  - Linear Regression -? Sum of squared errors
  - Neural Networks - ? Minimize cost function
- Most commonly used method is gradient descent

DATA FOLKZ
CATAPULT DATA LEADERS

GD - INTRODUCTION

- Y = x^3 + x^2 +3
- Lets relate this to it using calculus , ca
- Y= B0+B1(x)



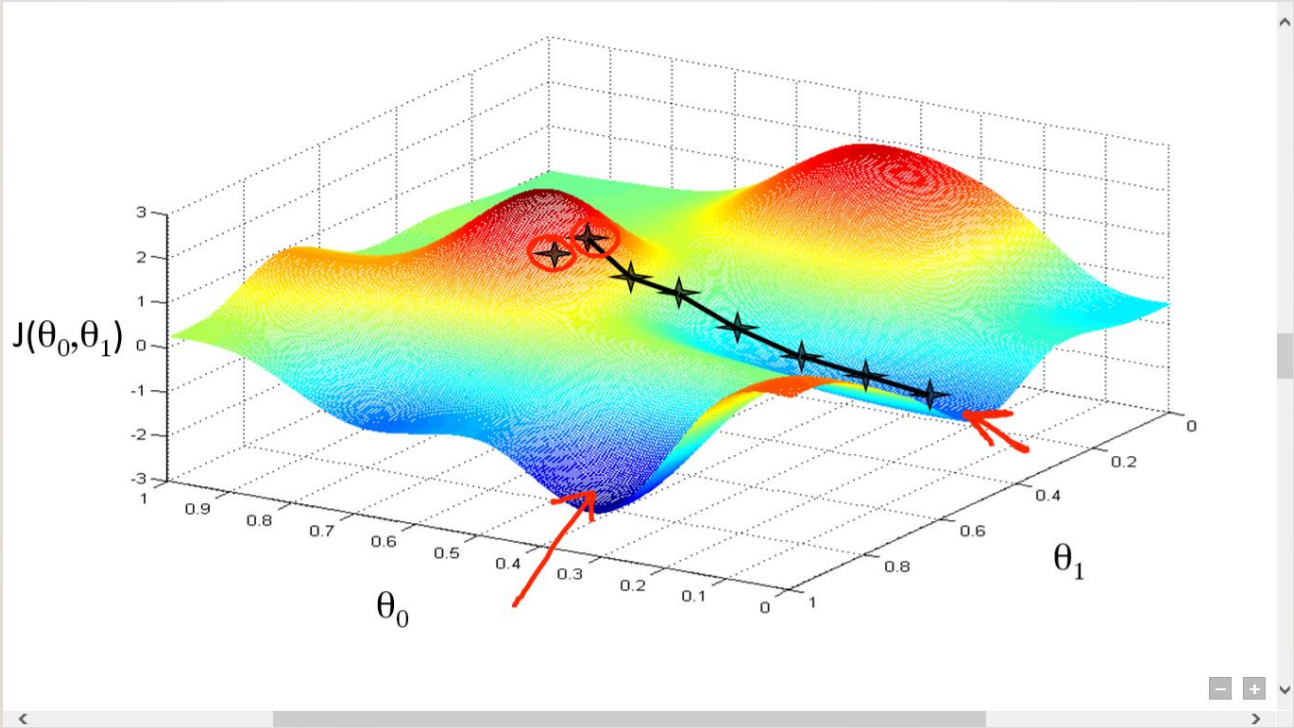
- Find the minimum:
  - Min f(x)
  - *f. : R^n → R*

In Calculus find the roots of the equation is not always easy, hence we use some maximization methods

GRADIENT

- The **gradient** is a fancy word for derivative, or the rate of change of a function
  - Points in the direction of greatest increase of a function
- The regular, plain-old derivative gives us the rate of change of a single variable, usually x. For example, dF/dx tells us how much the function F changes for a change in x. But if a function takes multiple variables, such as x and y, it will have multiple derivatives: the value of the function will change when we "wiggle" x (dF/dx) and when we wiggle y (dF/dy)

DATA FOLKZ
CATAPULT DATA LEADERS



- We introduce a model for algorithm:

Data $x_0 \in R^n$

Step 0: set $i = 0$

Step 1: if $\nabla f(x_i) = 0$ **stop**,

else, compute ***search direction*** $h_i \in R^n$

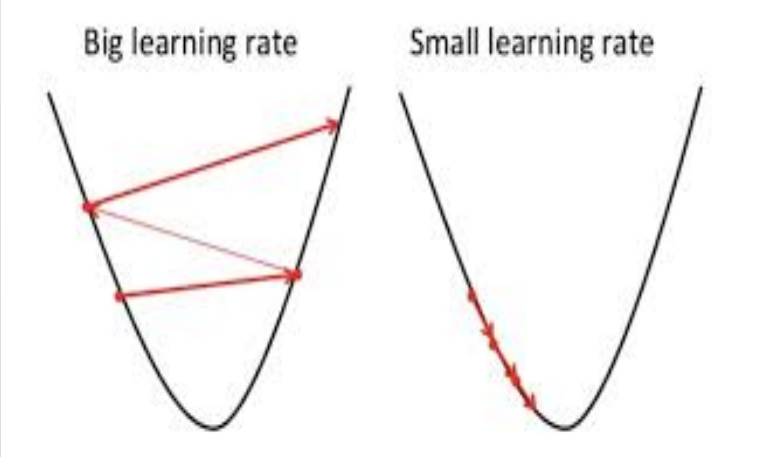
Step 2: compute the ***step-size*** $\lambda_i \in \arg\min_{\lambda \geq 0} f(x_i + \lambda \cdot h_i)$

Step 3: set $x_{i+1} = x_i + \lambda_i \cdot h_i$ go to step 1

DATA FOLKZ
CATAPULT DATA LEADERS

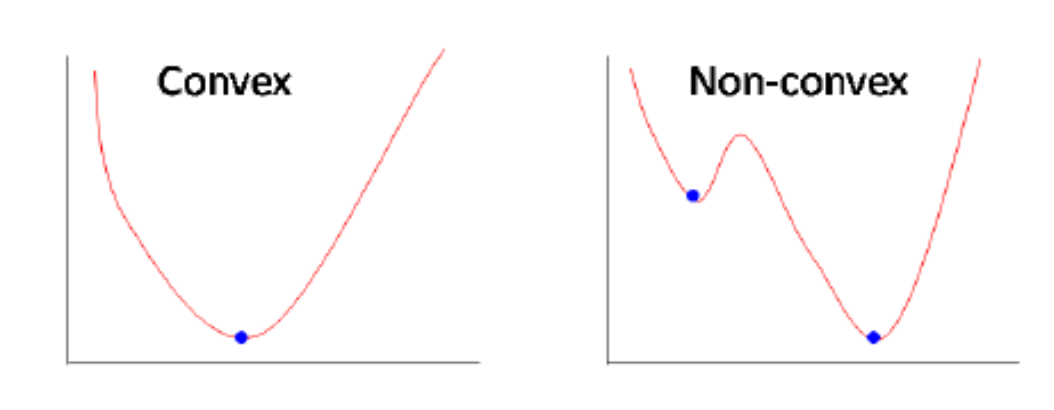
LEARNING RATE

- This is called step size, lets say I am climbing down, if very small it takes time, if its more I might miss the optimal minimum.



DATA FOLKZ
CATAPULT DATA LEADERS

WHEN DOES GD WORKS THE BETTER?



When you have multiple minimas , you arrive at different minimas, hence we mostly try to choose functions which has a convex shape

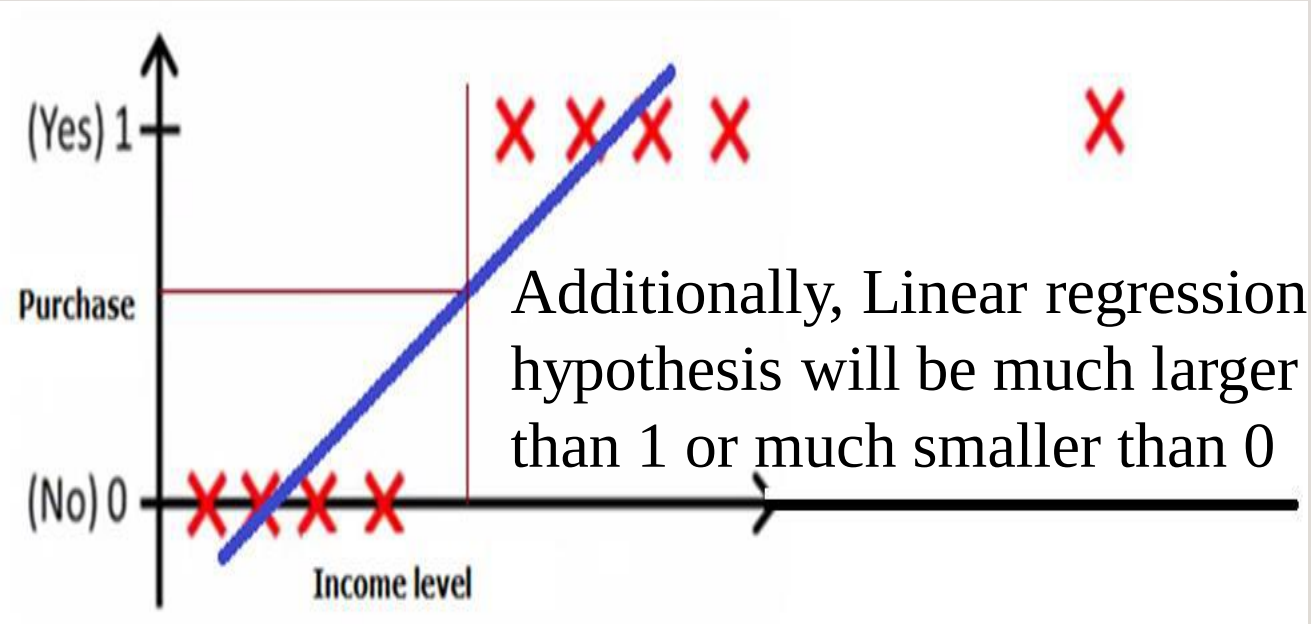
LOGISTIC REGRESSION

DATA FOLKZ
CATAPULT DATA LEADERS

HOW DIFFERENT IS FROM LINEAR?

DATA FOLKZ
CATAPULT DATA LEADERS

IT COULD FAIL

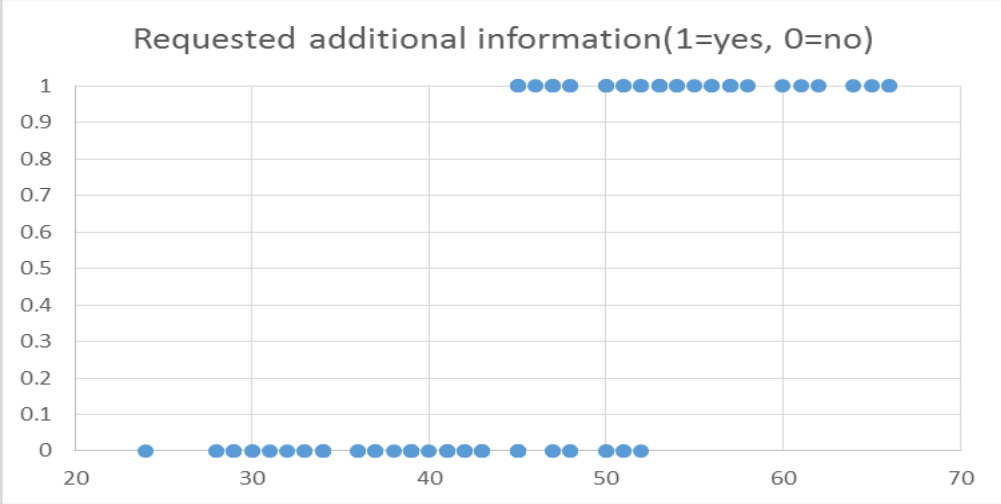


EXAMPLE

- An auto club mails a flier to its members offering to send more information regarding a supplemental health insurance plan if the member returns a brief enclosed form.
- Can a model be built to predict if a member will return the form or not?

DATA FOLKZ
CATAPULT DATA LEADERS

EXAMPLE



EXAMPLE

f(x) = p = 1/(1+e^- μ)

Where μ=β0+β1x1(also known as the systematic or the structural component or linear predictor).

- This is a logistic model. The function is also known as the inverse link function, which links the response with the systematic component.
  p is the probability that a club member fits into group 1 (returns the form; success; P(Y=1|X)).

LOGISTIC MODEL

- $$f(x) = p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

- Logistic model can be transformed into an odds ratio:

  S = odds ratio = p / (1-p)

ODDS RATIO

- If the probability of winning is 6/12, what are the odds of winning?
- If the odds of winning are 13:2, what is the probability of winning?
- If the odds of winning are 3:8, what is the probability of losing?
- If the probability of losing is 6/8, what are the odds of winning?

DATA FOLKZ
CATAPULT DATA LEADERS

LOGISTIC MODEL

$$S = Odds\ ratio = \frac{p}{1-p}$$

$$S = \frac{\frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}{1 - \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}} \qquad \therefore, S = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$$

Text Text

$$\ln(S) = \ln\left(e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

DATA FOLKZ
CATAPULT DATA LEADERS

LOGISTIC MODEL

Least Squares Vs MLE

- In Regression we minimized

$$SSE = \sum (y_i - \hat{y}_i)^2$$

- In Logistic Regression we maximize log likelihood instead

$$\text{Log Likelihood} = \sum [\, y_i \ln \hat{y}_i \; + \; (1 - y_i) \ln(1 - \hat{y}_i)]$$

- Log Likelihood is a convex function and hence finding optimal parameters is easier.

DATA FOLKZ
CATAPULT DATA LEADERS

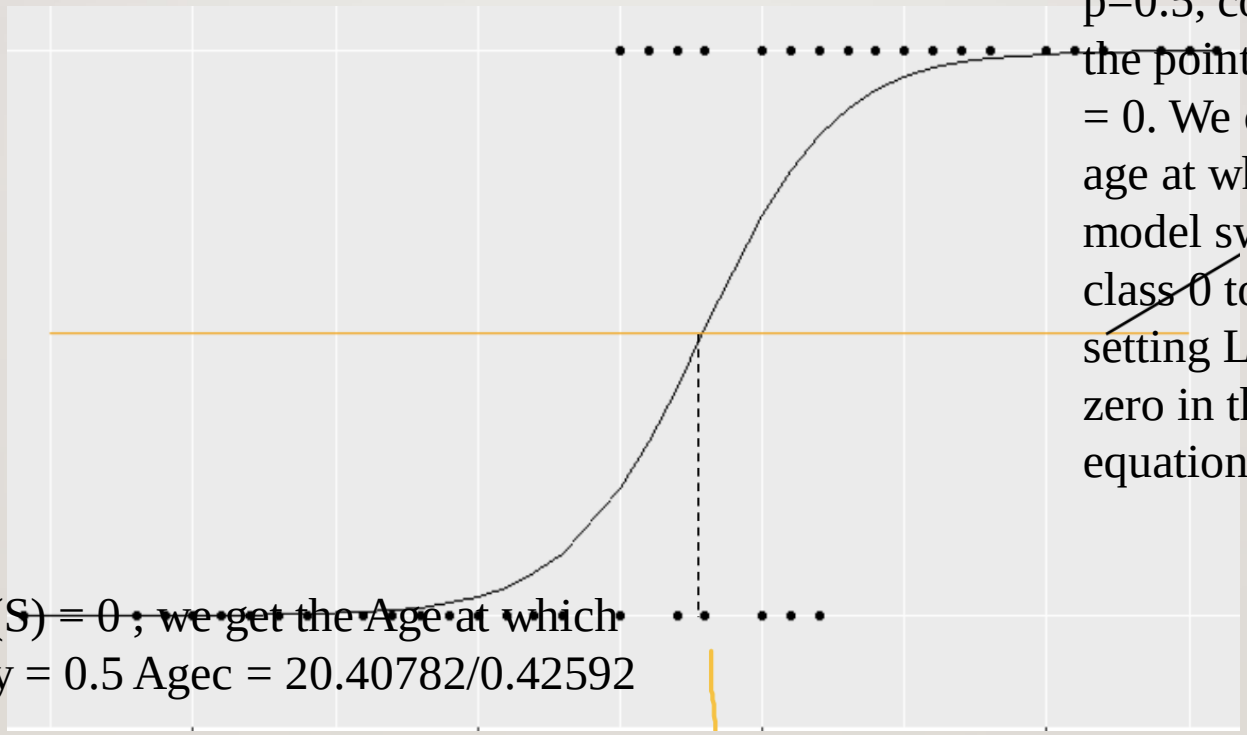
MAX LIKELIHOOD ESTIMATION

- Likelihood is reverse of probability
- In probability, we predict data based on known parameters
  - Eg (In binomial, we predict the probability with the number of trials etc.,)
- In likelihood, we predict parameters based on known data.

MLE

- Goal is to maximize likelihood
- We use calculus , or gradient descent for arriving at minimal
- ln S = −20.40782 + 0.42592 Age
- Suppose we want a probability that a 50-year old club member will return the form
- ln S = −20.40782 + 0.42592 ∗ 50 = 0.89 S = e 0.89 = 2.435
- The odds that a 50-year old returns the form are 2.435 to 1.

VISUALIZING THE FIT



The threshold of p=0.5, corresponds to the point where Ln(S) = 0. We can obtain the age at which the model switches from class 0 to class 1, by setting Ln(S) to be zero in the logistic equation.

Setting ln(S) = 0 ; we get the Age at which probability = 0.5 Agec = 20.40782/0.42592 = 47.9