COMP4388: **MACHINE LEARNING**

Logistic Regression

- Into classification
- Logistic Regression
- Handling Multiclasses

BIRZEIT UNIVERSITY

Dr. Radi Jarrar
Department of Computer Science

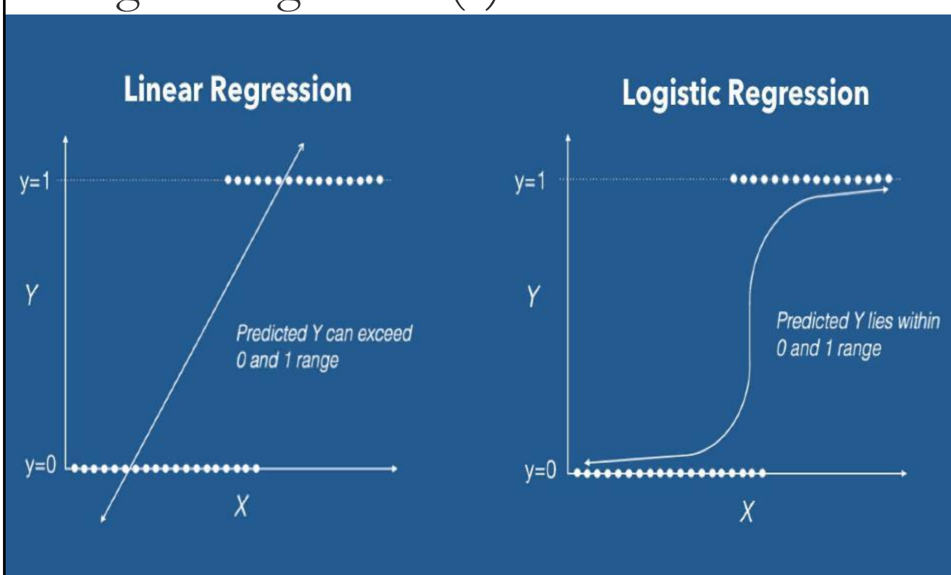
Regression Technique – Logistic

- A statistical model that uses a Logistic function to model a binary dependent variable
- Logistic regression is used to find the probability of event=Success and event=Failure
- It should be used when the dependent variable is binary (e.g., True/ False, Yes/ No)
- Problem of linear regression: Binary data is not normally distributed

Logistic Regression

- The core of the model is $h(x) = a^T x$, which combines the input variables linearly
- In linear regression, the output of the function h(x) is taken as the **real** value representing the output
- In linear classification, the output of the linear regression is thresholded to produce a **bounded** output of (-1/+1) which is appropriate for classification tasks

Logistic Regression (2)



Logistic Regression (3)

- Another possibility is to output a probability between 0 and 1
- It is similar to the previous models as the output is real (*as in regression*) but bounded (*as in classification*)
- Logistic Regression is a well-known classifier and widely used for binary classification problems

Logistic Regression (4)

- Linear classification uses hard threshold

$$h(x) = sign(a^T x)$$

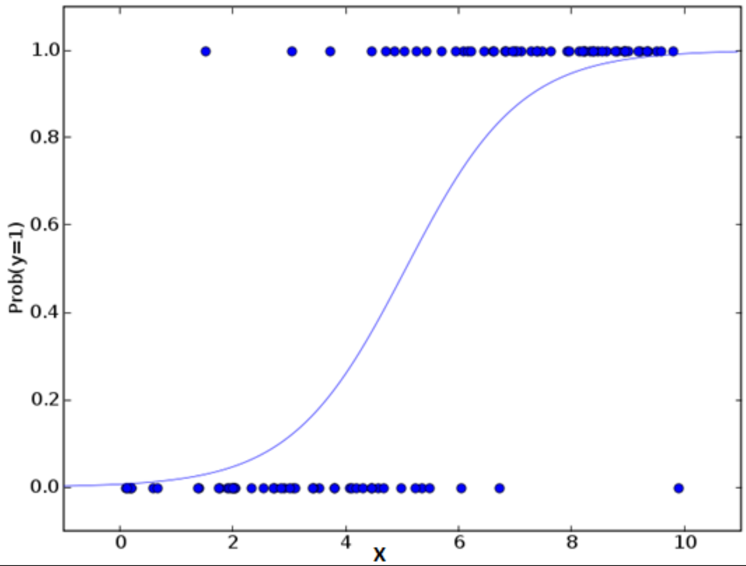
- Linear regression uses no threshold

$$h(x) = a^T x$$

- In logistic regression, a compromise of both models is made such that it restricts the output to the probability range [0, 1]
- This is done through the following model

$$h(x) = q(a^T x)$$

Logistic Regression (5)



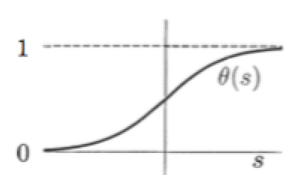
Logistic Regression (6)

$$h(x) = q(a^T x)$$

- In which q is the logistic function and its output is between 0 and 1

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{e^{\alpha_0+\alpha_1 x_1+\cdots+\alpha_n x_n}}{1+e^{\alpha_0+\alpha_1 x_1+\cdots+\alpha_n x_n}}$$

- The output is interpreted as the a probability for a binary event



Logistic Regression (7)

- The logistic function is a link function that is best suited for the binomial distribution
- The parameters are chosen to maximise the liklehood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression)

Logistic Regression (8)

- Linear classification deals with binary events but the difference is that logistic regression is allowed to be uncertain with intermediate values between 0 and 1 reflecting this uncertainty
- Logistic regression function is known as soft threshold
- It is also called the sigmoid function

Logistic Regression (9)

- It is widely used for classification problems
- No linear relationship required (as it applies a non-linear log transformation to the predicted odds ratio)
- Required a large sample size (Max likelihood estimates are less powerful with small sample size)

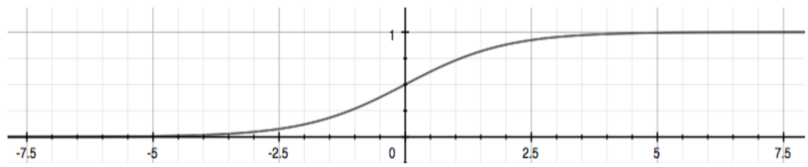
Hypothesis Representation

- The hypothesis representation, which is the function that we will use to represent a hypothesis when we have a classification problem
- Using a simple linear regression to approach a classification problem has the problem that predicting y might get larger than 1 or smaller than zero (given a value of x)

Hypothesis Representation (2)

- h(x) is modified to satisfy $0 \leq h(x) \leq 1$
- This is accomplished by plugging $\alpha^T x$ into the Logistic function

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



Hypothesis Representation (3)

- The sigmoid function maps any real value to the range (0, 1), which is more suited for classification
- Accordingly, h(x) is the estimated probability that y = 1 on input x

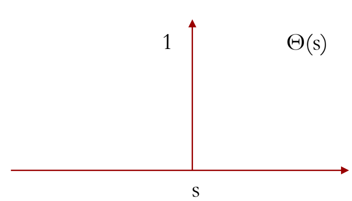
Hypothesis Representation (4)

- For example, x=CA-125 marker, y = 1 (if the tumor is malignant). If h(x) = 0.75, this means that the probability is 75% that the output is 1 (meaning, it is 75% that the tumor is malignant)
- Formally:

$$h(x) = p(y = 1 | x; \alpha)$$

Decision Boundary

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



- The sigmoid function slowly increases from zero to 1
- Suppose predict 'y=1' if h(x) ≥ 0.5

  predict y=0 if h(x) < 0.5

Decision Boundary (2)

- In order to get our discrete 0 or 1 classification, the output of the hypothesis function is translated as:

  $h(x) \geq 0.5 \rightarrow y=1$

  $h(x) < 0.5 \rightarrow y=0$

- Logistic function gives an output greater than or equal to zero when the input is greater than or equal to zero

$\theta(s) \geq 0.5$ when $s \geq 0$

Decision Boundary (3)

- So if the input to θ is $\alpha^T x$, then that means:

  $h(x) = \theta(\alpha^T x) \geq 0.5$ when $\alpha^T x \geq 0$

- Which means:

  $\alpha^T x \geq 0 \Rightarrow y=1$

  $\alpha^T x < 0 \Rightarrow y=0$

- Notes, when:

$s=0,\ e^0=1 \Rightarrow \theta(s)=0.5$

$s \rightarrow \infty, e^{-\infty} \rightarrow 0 \Rightarrow \theta(s)=1$

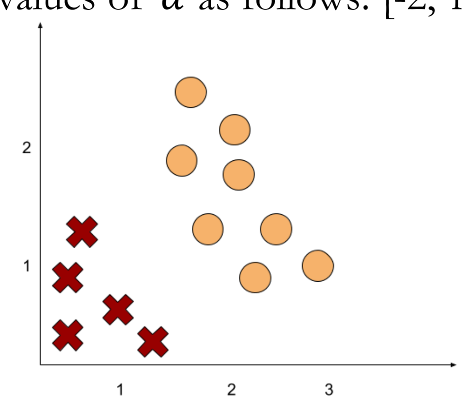
$s \rightarrow -\infty, e^{\infty} \rightarrow \infty \Rightarrow \theta(s)=0$

Decision Boundary (4)

- Decision boundary is the line that separates the area where y = 0 and where y = 1
- It is created by our hypothesis function
- How Logistic regression behaves with more than one feature?

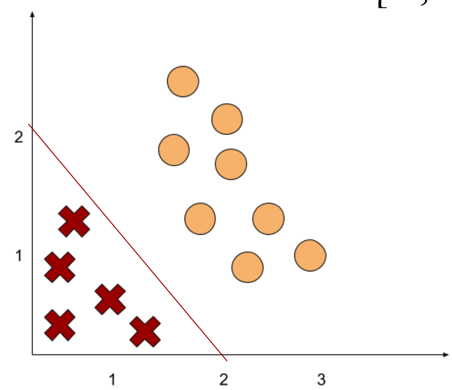
Decision Boundary (5)

- Assume there are two variables x_1 and x_2
- Accordingly, $h(x) = \theta(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$
- Assume GD found the values of α as follows: [-2, 1, 1]



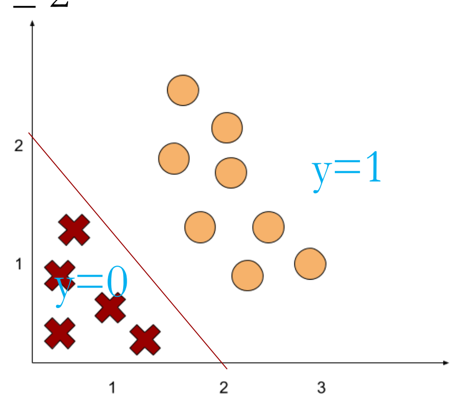
Decision Boundary (6)

- Assume there are two variables x_1 and x_2
- Accordingly, $h(x) = \theta(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$
- Assume GD found the values of α as follows: [-2, 1, 1]
- Predict 'y=1' if
- $\alpha^T x \geq 0$, means $-2 + x_1 + x_2 \geq 0$



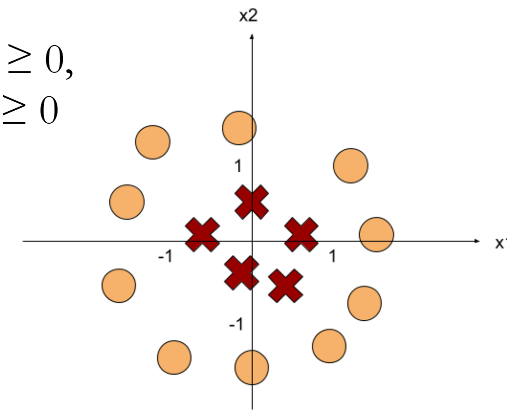
Decision Boundary (7)

- Predict 'y=1' if
- $\alpha^T x \geq 0$, means $-2 + x_1 + x_2 \geq 0$
- This also means $x_1 + x_2 \geq 2$
- Such that $x_1 + x_2 = 2$
- Which is the equation of a straight line
- Y=0 when $x_1 + x_2 < 2$



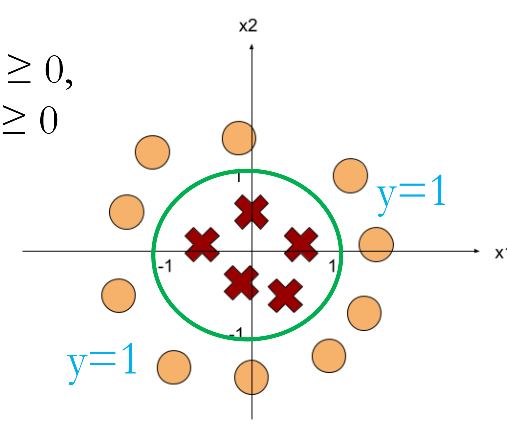
Non-linear decision boundary

- $h(x) = \theta(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1^2 + \alpha_4 x_2^2)$
- Assume GD found the values of α as follows: [-1, 0, 0, 1, 1]
- Predict 'y=1' if $\alpha^T x \geq 0$, means $-1 + x_1^2 + x_2^2 \geq 0$
- Which also means $x_1^2 + x_2^2 \geq 1$ (which is the equation of a circle)



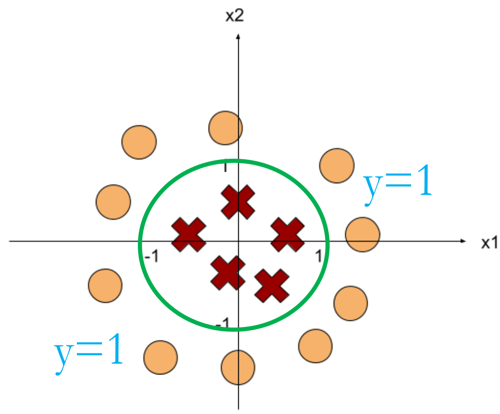
Non-linear decision boundary

- $h(x) = \theta(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1^2 + \alpha_4 x_2^2)$
- Assume GD found the values of α as follows: [-1, 0, 0, 1, 1]
- Predict 'y=1' if $\alpha^T x \geq 0$, means $-1 + x_1^2 + x_2^2 \geq 0$
- Which also means $x_1^2 + x_2^2 \geq 1$ (which is the equation of a circle)



Non-linear decision boundary (2)

- The higher order polynomials created more complex decision boundaries to separate the positive and negative examples



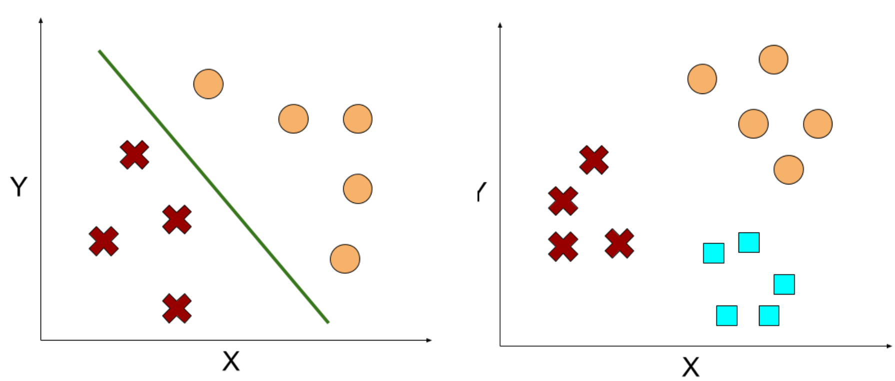
MULTICLASS CLASSIFICATION

One-against-all

Multiclass classification

- Binary classification problems are when there are only two classes (0 or 1), (-1 or 1)
- In Multiclass classification, there are more than two classes (i.e., classifying images into semantic categories, classifying music according to there genres, classifying emails to different set of labels or folders, …)

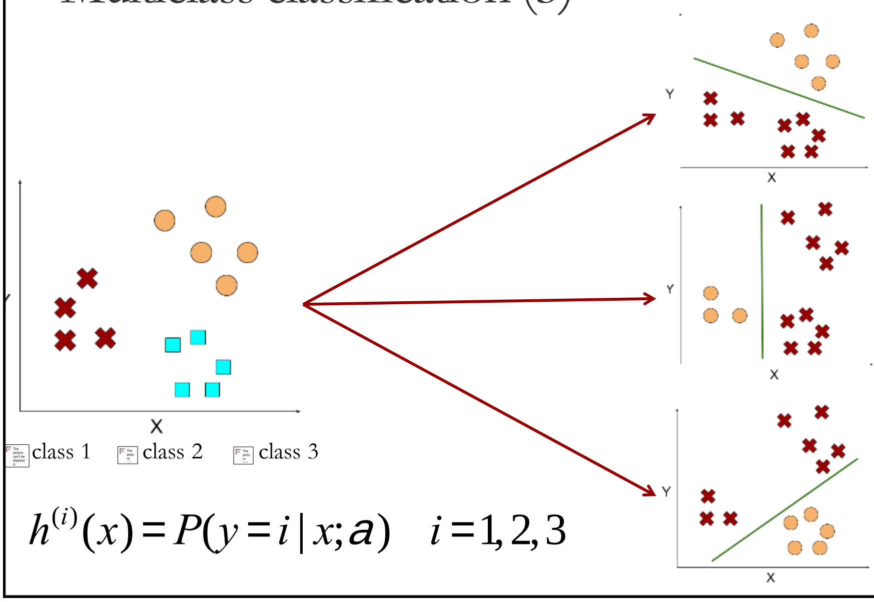
Multiclass classification (2)



Binary classification

Multiclass classification

Multiclass classification (3)



$$h^{(i)}(x) = P(y = i \mid x; a) \quad i = 1, 2, 3$$

Multiclass classification (4)

- Train a logistic regression classifier $h^{(i)}(x)$ for each class i to predict the probability that y=i
- On a new input instance x, classify it with the class target i that maximises the prediction

$$\max_i h^{(i)}(x)$$