Week 6

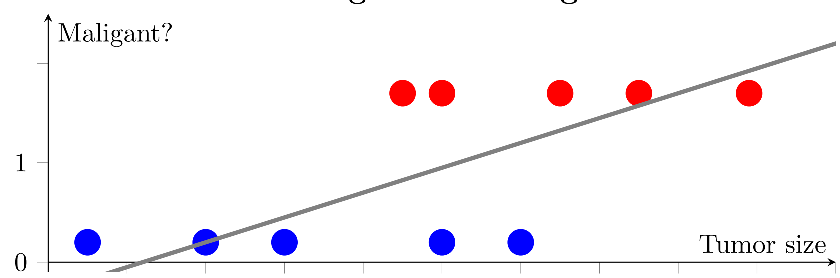
August 18, 2021

Classification

- Y can only have discrete values.
- For example: 0 = negative class (absence of something) and 1 = positive class (presence of something).
- Email $->$ spam/not spam?
- Online transactions $->$ fraudulent?
- Tumor $->$ Malignant/benign?

Let's go back to the cancer example from the Week 1 and try to apply linear regression:

Define breast cancer as malignant or benign based on tumour size



We see that it wasn't the best idea. Of course, we could attempt another approach to find a straight line that would better separate the points, but a straight line isn't our sole choice. There are more appropriate functions for that job.

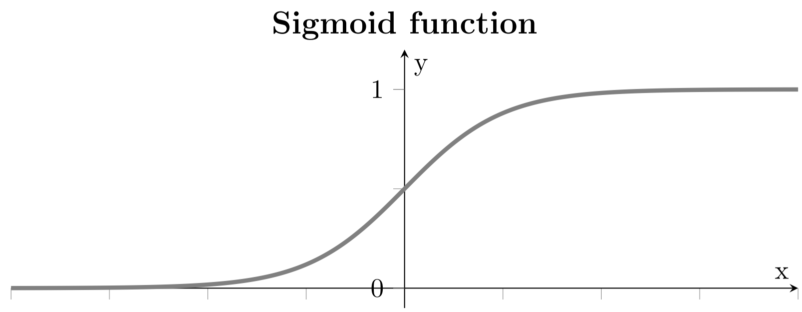
Hypothesis representation

- We want our classifier to output values between 0 and 1.
- For classification hypothesis representation we have: $h_\theta(x) = g((\theta^T x))$.
- $g(z)$ is called the sigmoid function, or the logistic function.

$$g(z) = \frac{1}{1 + e^{-z}}$$

- If we combine these equations we can write out the hypothesis as:

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



When our hypothesis ($h_\theta(x)$) outputs a number, we treat that value as the estimated probability that $y = 1$ on input x.

$$h_\theta(x) = P(y = 1 | x \; ; \; \theta)$$

Example:

$h_\theta(x) = 0.7$ and

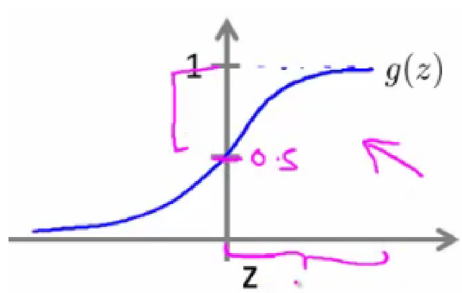
$$X = \begin{bmatrix} 1 \\ tumourSize \end{bmatrix}$$

Informs a patient that a tumor has a 70% likelihood of being malignant.

Decision boundary

One way of using the sigmoid function is:

- When the probability of y being 1 is greater than 0.5 then we can predict y = 1.
- Else we predict y = 0.



- The hypothesis predicts $y = 1$ when $\theta^T x >= 0$.
- When $\theta^T x <= 0$ then the hypothesis predicts y = 0.

Example

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

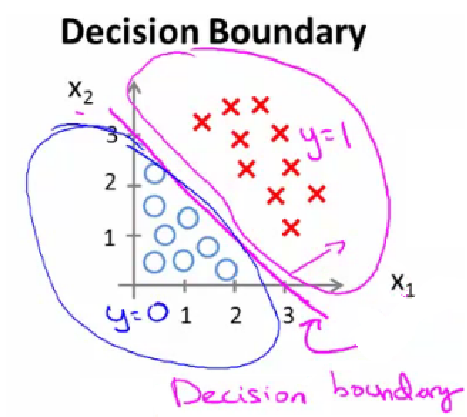
$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

We predict $y = 1$ if:

$$-3x_0 + 1x_1 + 1x_2 \geq 0$$
$$-3 + x_1 + x_2 \geq 0$$

As a result, the straight line equation is as follows:

$$x_2 = -x_1 + 3$$



- Blue = false
- Magenta = true
- Line = decision boundary

Non-linear decision boundaries

Get logistic regression to fit a complex non-linear data set.

Example

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

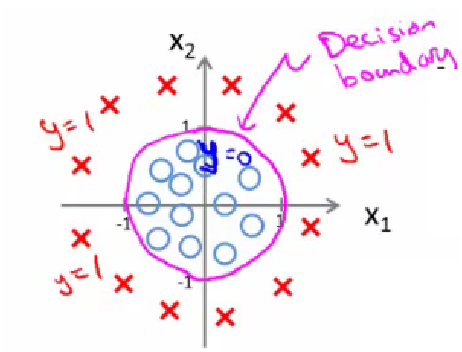
We predict $y = 1$ if:

$$-1 + x_1^2 + x_2^2 \geq 0$$
$$x_1^2 + x_2^2 \geq 1$$

As a result, the circle equation is as follows:

$$x_1^2 + x_2^2 = 1$$

This gives us a circle with a radius of 1 around 0.



Cost function for logistic regression

- Fit θ parameters/
- Define the optimization object for the cost function we use the fit the parameters.

Training set of m training examples:

$$\{(x^{(1)}, y^{(1)}), (x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})\}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ ... \\ x_n \end{bmatrix}$$

$$x_0 = 1, \quad y \in \{0, 1\}$$

Linear regression uses the following function to determine θ:

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

We define "cost()" as:

$$cost(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$$

We can now redefine $J(\theta)$ as:

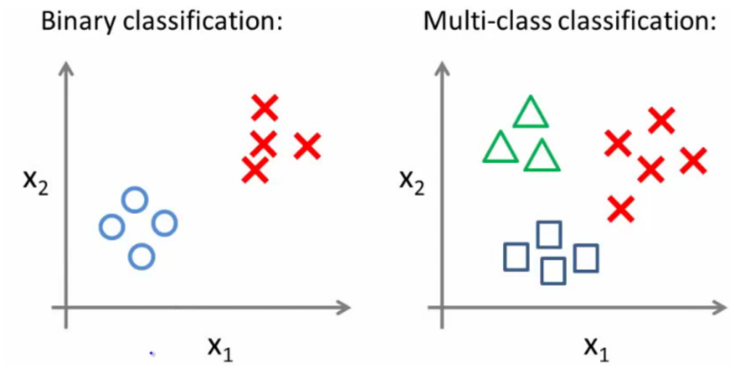
$$J(\theta) = \frac{1}{2} \sum_{m}^{i=1} cost(h_\theta(x^{(i)}), y^{(i)})$$

- This is the cost you want the learning algorithm to pay if the outcome is $h_\theta(x)$ but the actual outcome is y.
- This function is a non-convex function for parameter optimization when used for logistic regression.
- If you take $h_\theta(x)$ and plug it into the Cost() function, and them plug the Cost() function into $J(\theta)$ and plot $J(\theta)$ we find many local optimum.

$$cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & if\ y = 1 \\ -log(1 - h_\theta(x)) & if\ y = 0 \end{cases}$$

Multiclass classification problems

Getting logistic regression for multiclass classification using one vs. all.



Split the training set into three separate binary classification problems.

- Triangle (1) vs crosses and squares (0) $h_\theta^{(1)}(x)$.
- Crosses (1) vs triangle and square (0) $h_\theta^{(2)}(x)$.
- Square (1) vs crosses and square (0) $h_\theta^{(3)}(x)$.