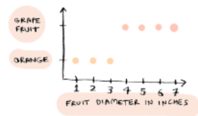
Logistic Regression

- Name is somewhat misleading. Really a technique for classification, not regression.
  - "Regression" comes from fact that we fit a linear model to the feature space.
- Involves a more probabilistic view of classification.

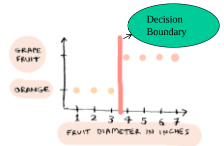
Suppose you are trying to classify a new piece of fruit as an orange or a grapefruit.

- First you plot your data on a graph. Here's your existing data:



i.e. anything that is less than 4 inches in diameter is an orange, anything greater than that is a grapefruit.

- Now you can draw a line in this image. Anything to the left of the line is an orange, and anything to the right is a grapefruit.



Prediction Function

For Regression Problem:

X	y
1000	$200k
2000	$250k
4000	$300k

i.e. **home size -> price of home**.

For Logistic Regression:

FRUIT DIAMETER IN INCHES

X	y
1	Ø
2	Ø
3	Ø
4	1
5	1

Ø = NOT A GRAPEFRUIT
1 = GRAPEFRUIT

i.e. **fruit diameter -> is it a grapefruit**

For logistic regression, we predict a probability, like "there's a 90% chance that this is a grapefruit".

For linear regression had the hypothesis as:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Once you had the optimal values for theta_0 and theta_1, the prediction part was easy. We use almost the same formula for logistic regression, with one change:

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

SIGMOID FUNCTION

As we need to predict percentage measure of likeliness. **The sigmoid function constrains the result to between 0 and 1**. So a result of .95 would mean we are 95% certain of something.

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

ACTUAL VALUE
1 = IS A GRAPEFRUIT

X	y	PREDICTED
1	Ø	Ø.6%
2	Ø	4.7%
3	Ø	26.9%
4	1	73.1%
5	1	95.2%

PERCENT CHANCE THAT THIS IS A GRAPEFRUIT

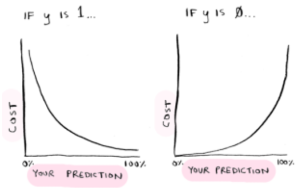
The sigmoid function smoothes out the answers, instead of just splitting out a yes or a no:

Cost Function

- In linear Regression the cost was based on how far off the line was from our data points.
- In logistic regression the cost is calculated based on how far off your probability was.
- Eg: If we say "There's a 95% chance that this is a grapefruit", but it turns out to be an orange, we should get penalized heavily (i.e. the cost should be higher). But if we say "There's a 55% chance that this is a grapefruit" and it turns to be an orange, that should be a lower penalty (i.e. the cost should be lower).

Cost Function

- y is the actual result. So if y = 1 i.e. it was a grapefruit, and you predicted 1 (i.e. 100% probability it is a grapefruit), there's no penalty. But if you predicted 0 (0% probability it is a grapefruit), then you get penalized heavily.



Cost Function

- The more wrong you are, the more you get penalized. We use the log function for this.
- Cost function is, hence,

$$-\frac{1}{m}\sum_{i=1}^{m} \begin{cases} -\log(h_\theta(x_i)) & \text{if } y_i = 1 \\ -\log(1-h_\theta(x_i)) & \text{if } y_i = \emptyset \end{cases}$$

- More refined:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \underbrace{y_i \cdot \log(h_\theta(x_i))}_{\text{THIS EVALUATES TO ZERO IF } y_i = \emptyset} + \underbrace{(1-y_i)\cdot \log(1-h_\theta(x_i))}_{\text{THIS EVALUATES TO ZERO IF } y_i = 1}$$

- It can be optimized using gradient descent too.

Multiclass Classification

- When we have to classify more than two classes, we need more decision boundaries.

(O = orange, G = grapefruit, A = apple).

One-Vs-All Classification

- We just concentrate on one class at a time and treat the rest as one separate class.

- So for I fig. either the fruit IS a grapefruit, or ISN'T a grapefruit (oranges and apples temporarily get lumped together). You do this classification and get a number, lets say 25%. So there's a 25% chance this is a grapefruit, and *there's a 75% chance this is something other than a grapefruit*.

Nonlinear Regression: Intuition

- Suppose the data looks like this:

- There's no straight line that you can use as your decision boundary. You need a non-linear boundary, like this:

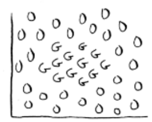
Nonlinear Regression: Intuition

- It uses a nonlinear hypothesis function, like:

$$h_\theta(x) = g\left(\theta_0 + \theta_1 x + \theta_2 (x-\theta_3)^2\right)$$

Where g(x) is a nonlinear function

- Linear regression is not good for something like this:



Nonlinear Regression

- A form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables.

- Nonlinear regression is characterized by the fact that the prediction equation depends nonlinearly on one or more unknown parameters.

$$Y_i = f(x_i, \theta) + \varepsilon_i$$

f is a known function of the covariate vector $x_i = (x_{i1}, x_{i2}, \ldots\ldots\ldots, x_{ik})$ and the parameter vector $\theta = (\theta_1, \theta_2, \ldots\ldots\ldots \theta_p)$, ε_i are random errors

Nonlinear Regression

A regression model is called “non linear” if the derivative of the model depends on one or more parameters

- Linear Model : Y = a + b X
- Linear Model : Y = a + b X + c X^2
- Non Linear Model : Y = a + b^2 X

Example

$$y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$$

is a linear model because $\partial y / \partial \beta_i$, $(i = 1,2,3)$ are independent of the parameters β_i, $(i = 1,2,3)$. On the other hand,

$$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X + \varepsilon$$

is a nonlinear model because $\partial y / \partial \beta_1 = 2\beta_1 X_1$ depends on β_1 although $\partial y / \partial \beta_2$ and $\partial y / \partial \beta_3$ are independent of any of the β_1, β_2 or β_3.