Review of Regression components

- Hypothesis: $H(X) = WX$
- Cost: $cost(W) = \frac{1}{m} \Sigma (WX - y)^2$
- Gradient Descent: $W := W - \alpha \frac{\partial}{\partial W} cost(W)$

Classification

Recall:

- Regression used to predict metric (e.g 0~100)
- classification used to predict binary result (e.g pass or fail)

Examples of classification:

- spam detection
- Facebook feed: show or hide
- Credit card Fraudulent Transaction detection

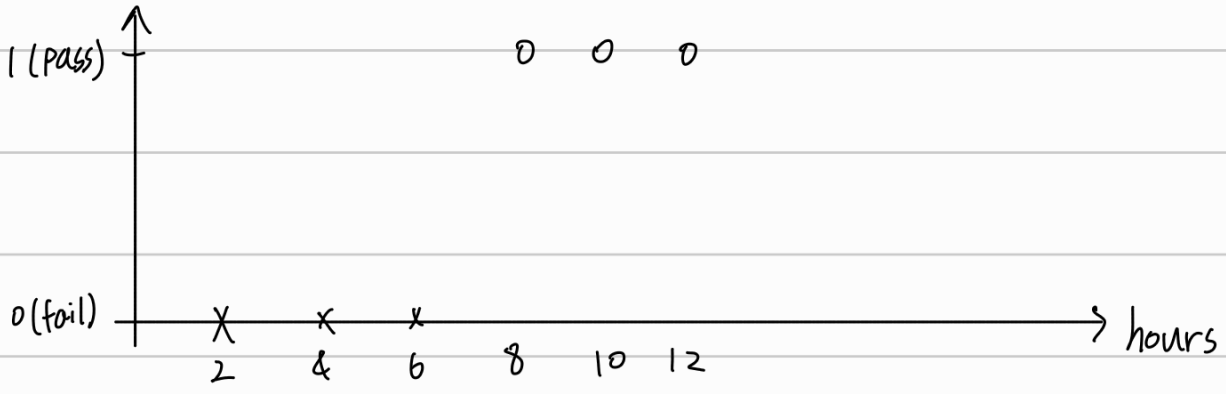
0, 1 Encoding

- Binary result ⇒ label w/ 0 or 1

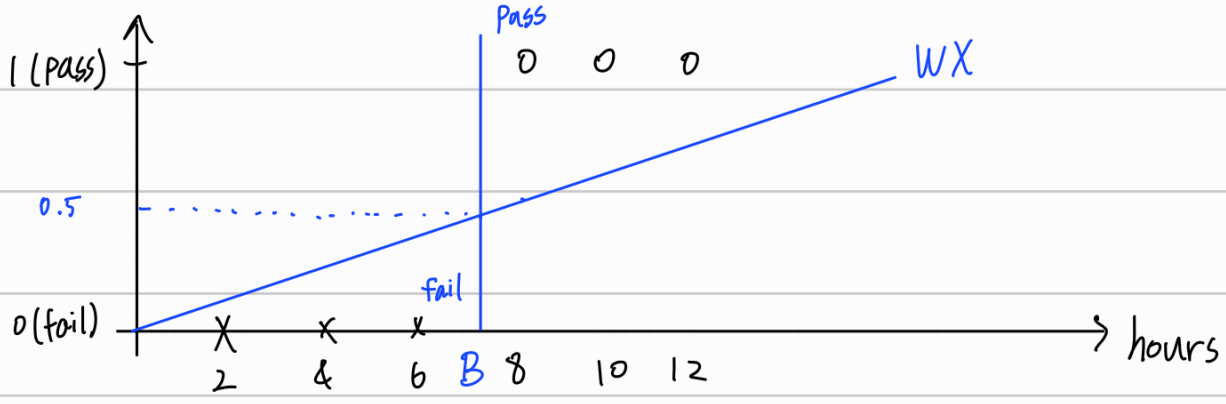
  Ex. spam (1) Ham (0) for spam detection

- Plotting looks like a binary step function

Ex. Pass (1) / Fail (0) based on study hours

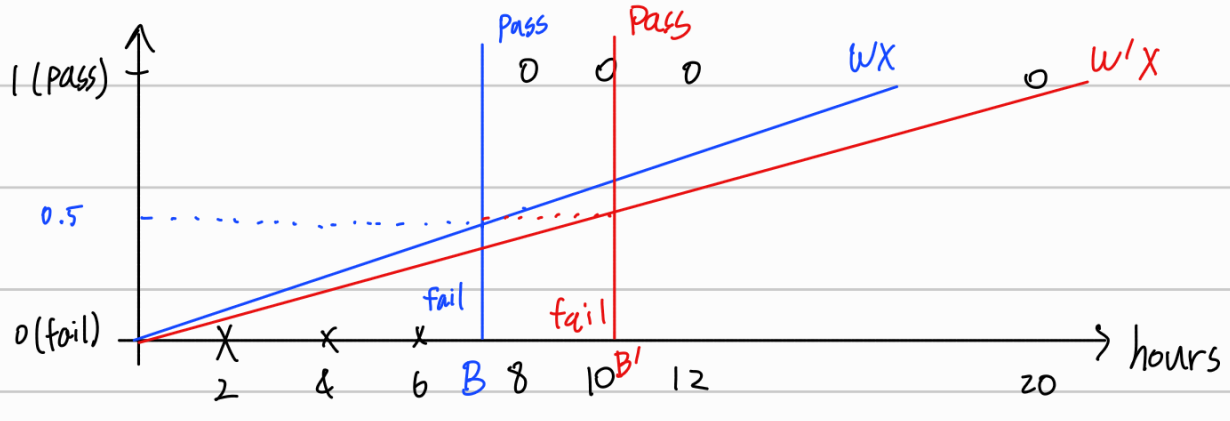


Can we use linear regression here?



The initial idea would be to construct a hypothesis WX drawn in blue. And then we find the "0.5" mark based on WX and determine P/F based on whether the x-input is less than or larger than the x-coord at the 0.5 mark (labelled B)

However, certain challenges arise:



- After introducing a new data point in our training data, the new optimized Hypothesis $W'X$ may shift the 0.5 mark (as seen w/ B'), causing previous data points that were labelled pass to be now labelled fail. (like the 8 and 10 hour data points above)

- Additionally, b/c we know Y is 0 or 1, our Hypothesis $H(x) = Wx + b$ can result in values larger than 1 or less than 0.

Ex. $W = 0.5$, $b = 0$. $H(x) = 0.5x$

new $x_i = 100 \Rightarrow H(x_i) = 50 >> 1$

↳ too large

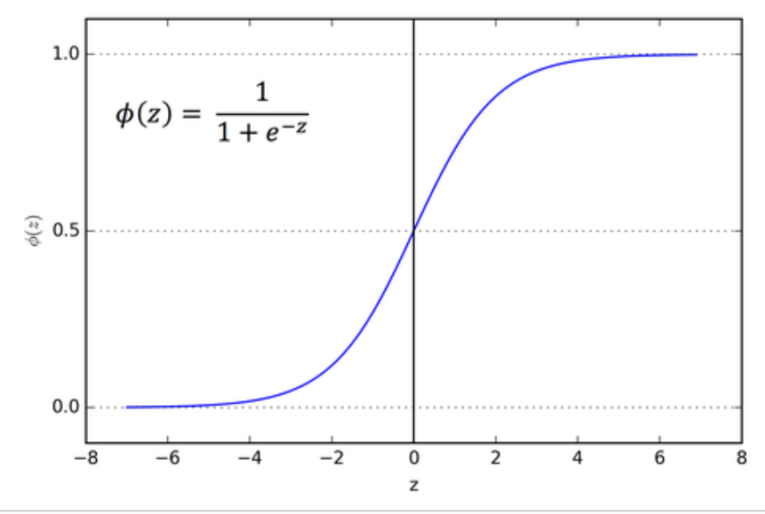
Solution: find a better function to construct a model for classification.

- Such a function needs to be bound in y-value btwn 0 and 1

∴ Logistic / Sigmoid function

↳ curved in 2 directions, like letter "S"

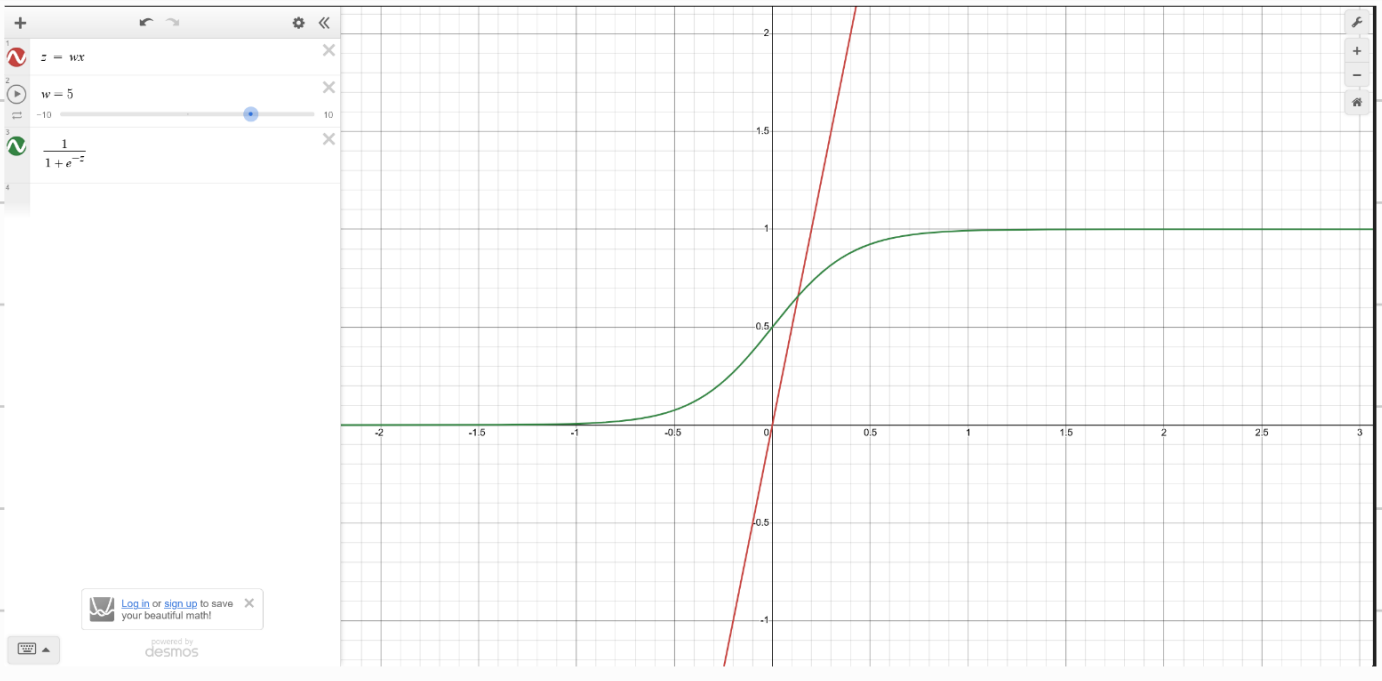
$$g(z) = \frac{1}{(1 + e^{-z})}$$

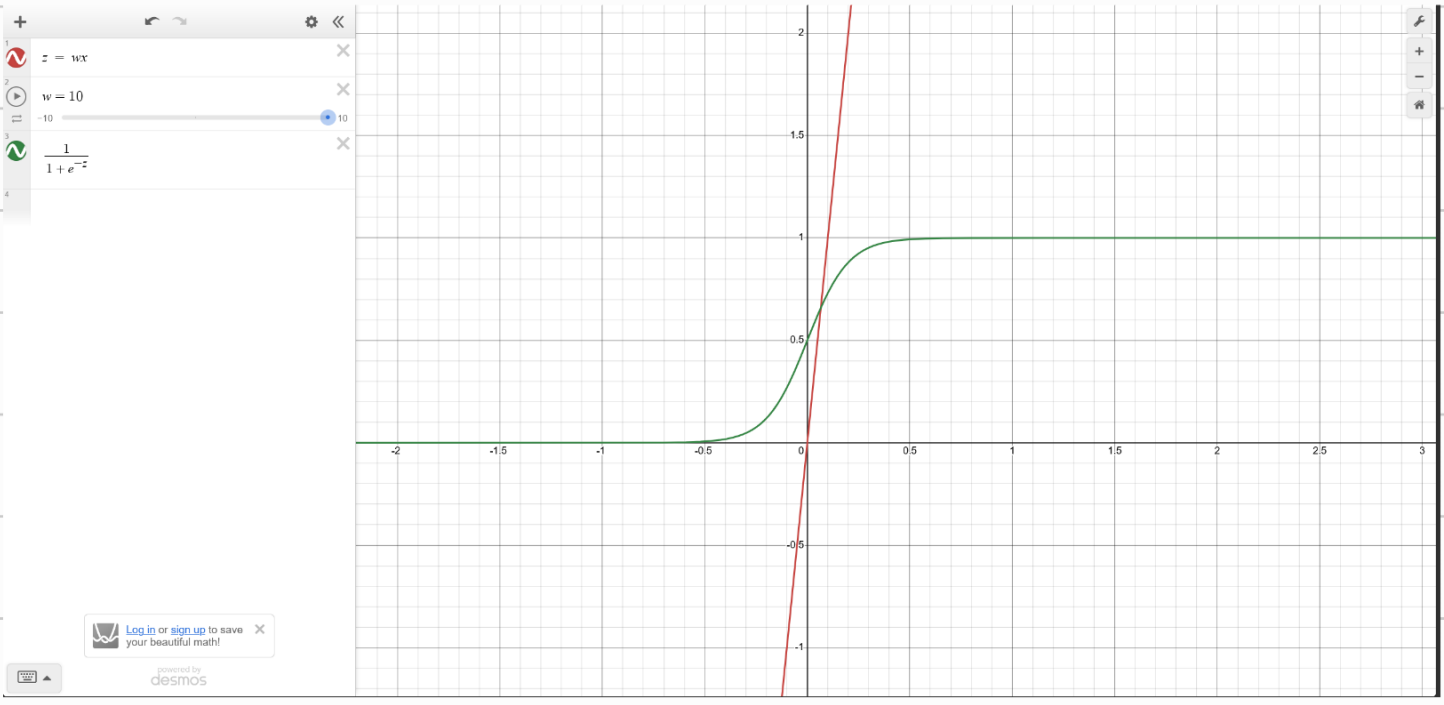


- For optimal results
  - Set $z = wx$
  - Set $H(x) = \frac{1}{1+e^{-z}}$

Why not just set $H(x) = \frac{1}{1+e^{-x}}$?

A: Setting weight (w) higher makes the sigmoid curve steeper, making "gray area" inputs easier to classify (x-values close to the 0.5 x-coord)





Logistic Classification Hypothesis

The new hypothesis is now constructed using the sigmoid function.

$$H(X) = \frac{1}{1 + e^{-(WX)}} \quad \text{or} \quad H(X) = \frac{1}{1 + e^{-(W^T X)}}$$

why are there 2 hypotheses?

A: If W has dimensions $m \times n$ and X has dimensions $m \times l$ for example, WX won't work (recall matrix multiplication $(a \times b) \cdot (b \times c) = (a \times c)$). Therefore, to match the dimensions, transpose W.