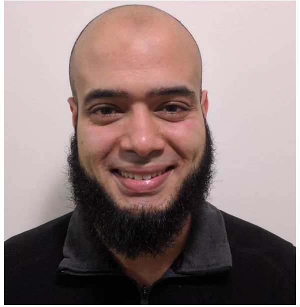
Machine Learning

Background
Sigmoid Function

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

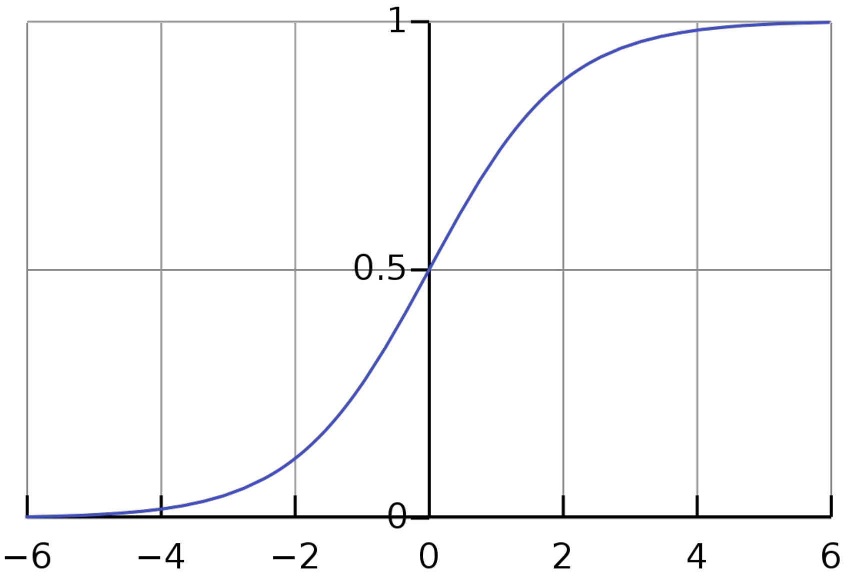


Sigmoid Function

- A sigmoid function is a **continuous** S-shaped function that maps (**squash**) any real value ([-OO, OO]) into range [0-1]
    - S(0) = 0.5
    - S(5) = 0.9933
    - S(10+) = almost one

$$S(x) = \frac{1}{1 + e^{-x}}$$

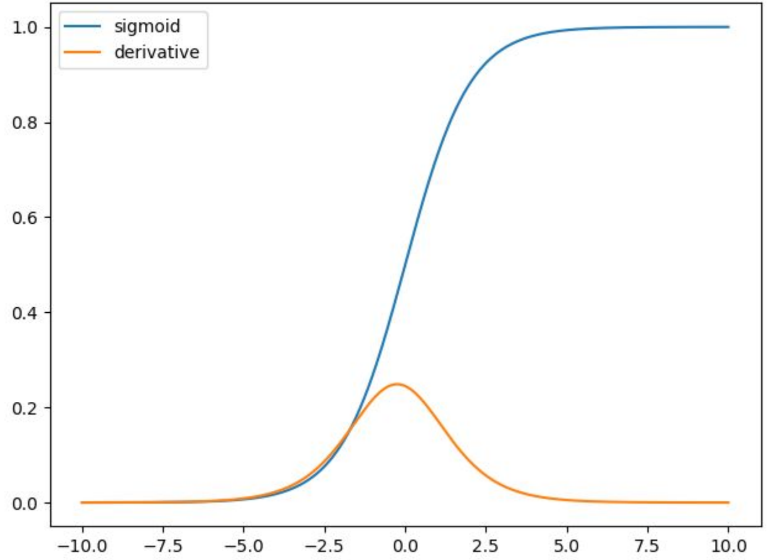
```
def sig(x):
 return 1/(1 + np.exp(-x))
```



Sigmoid Derivative

- The derivative of sigmoid function S(x) is just S(x) (1-S(x))
  - E.g. derivative at x = 0: is 0.5 x 0.5 = **0.25**
- This implies if we calculated S(x), then we **simply** can compute its derivative
- Sigmoid is *monotonic* but
  it's derivative is not
- Sigmoid has a *non-negative derivative*
  at each point and
  exactly one inflection point (x=0)

Img src



Sigmoid Derivative Proof

- Try to compute the derivative by yourself
  - Just math skills
- Then compare with this image (src)
  - Source also has a 2nd way
- *You might be asked such a question in interviews*

$$
\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] = \frac{d}{dx}(1+e^{-x})^{-1} \\
&= -1 * (1+e^{-x})^{-2}(-e^{-x}) \\
&= \frac{-e^{-x}}{-(1+e^{-x})^2} \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}}\,\frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}}\,\frac{e^{-x}+(1-1)}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}}\,\frac{(1+e^{-x})-1}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}}\left[\frac{(1+e^{-x})}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right] \\
&= \frac{1}{1+e^{-x}}\left[1 - \frac{1}{1+e^{-x}}\right] \\
&= \sigma(x)(1-\sigma(x))
\end{aligned}
$$

Sigmoid as an Activation Function

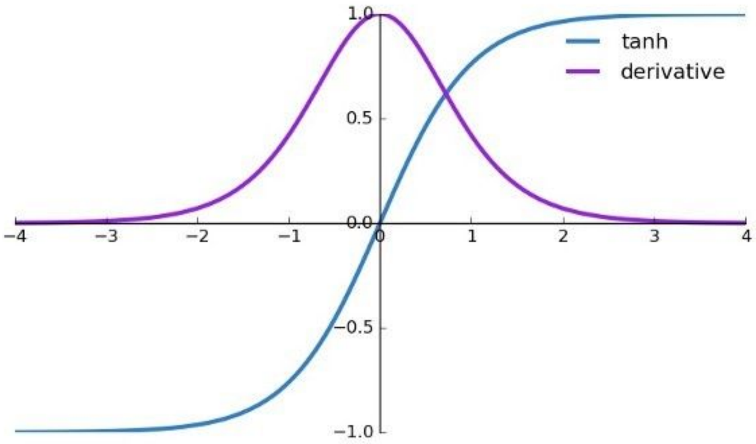
- This is a **note for future** for you
- In deep learning, there is an issue called vanishing gradients
  - E.g. the network is so deep and multiplying many small values = 0
  - The derivative of sigmoid is 0 for large inputs (10+)
- Nowadays, 95% we don’t use sigmoid or tanh as activations
  - Sigmoid is still there in the output layer for binary classification
  - RELU activation and its variants are much more stable
- Still there could be some scenarios where Sigmoid is used as an activation
  - For example, in LSTM network
  - Is being symmetric useful? No experience

Sigmoid output as a probability

- We will see this in coming lectures
- Given that the output of sigmoid is in range [0-1]
- We can **interpret** this output as a probability
- In terms of machine learning, we can define models (e.g. Neural Network), that estimates the **conditional probability** of an event
- P(dog | image) = 0.8
- **P** here is e.g. a deep neural network that takes input image and output its probability being a **dog**

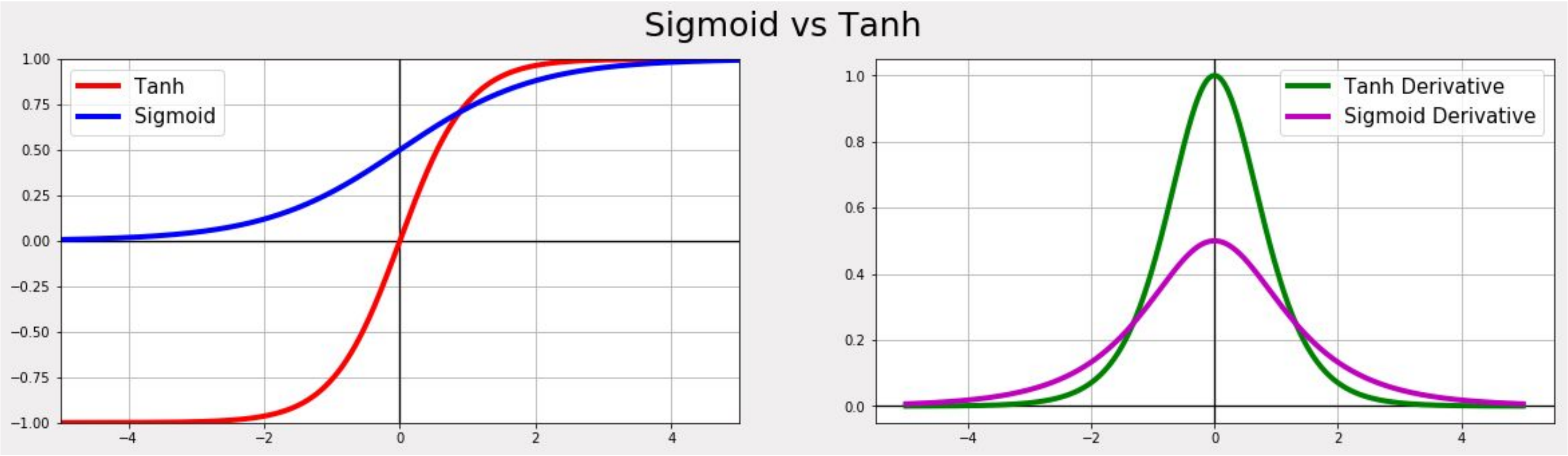
Tanh

- The hyperbolic tangent function, or tanh, is defined as: $$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$
- Tanh function is a rescaled sigmoid function
  - tanh(x) =2 sigmoid (2x) − 1
  - Try to prove
- The output of tanh is [-1, 1]
  - output **negative** values
  - output is **symmetric** around zero
- Its derivative is $1 - \tanh^2(x)$
- In GAN, if input images are scaled [-1, 1], you must use tanh as output
  - Otherwise sigmoid



Sigmoid vs Tanh

- Tanh has **stronger gradient** than sigmoid
  - f(0)’ = 0.25 vs 1
- Tanh has **steeper gradients** around zero ⇒ faster training
- Again, these notes are mainly useful for classical NN, rarely for deep learning



src

Tanh to Sigmoid Proof

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$= \frac{e^x + e^{-x} - 2e^{-x}}{e^x + e^{-x}}$$

$$= 1 + \frac{-2e^{-x}}{e^x + e^{-x}}$$

$$= 1 - \frac{2}{e^{2x} + 1}$$

$$\begin{aligned} \tanh(x) = 1 - \frac{2}{e^{2x}+1} &= 1 - 2\sigma(-2x) \\ &= 1 - 2(1 - \sigma(2x)) \\ &= 1 - 2 + 2\sigma(2x) \\ &= 2\sigma(2x) - 1 \end{aligned}$$

src

Logistic Function

- A more flexible version from the sigmoid (*but we typically don't use*)
  - x0: the x value of the function's midpoint;
  - k for the steepness of the curve
  - L affects the range of values to be in [-L to L], like scaling
- Sigmoid is called the standard logistic function (K=L=1, x0=0)
- Optional Homework: visualize the function for the different parameters to explore the different curves
  - Try: L {1, 2, 3}, K = {0.5, 1, 2}, x0 {0, 5, -8}
  - You may use Wolfram Alpha

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

Question!

- If a = sqrt(b): inverse it and find b = ?
- Now, inverse the sigmoid. Given s(x), find x

$$S(x) = \frac{1}{1 + e^{-x}}$$

Sigmoid vs Logit Function

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

- The logit function is the **inverse of the sigmoid**
  - Transforms [0, 1] to [-OO, OO]
  - Think of range [0, 1] as probability
- p is a probability
  - p/(1 − p) is called odds
    - ratio of the number of events that **produce** that outcome to the number that do **not**
  - Logit is called the logarithm of the odds (log-odds)
- There are a lot of history behind this function
  - You will hear this word a lot in **deep learning**
- Future: The logit in logistic regression is a special case of a link function in a **generalized linear model (GLM)**
  - it is the canonical link function for the Bernoulli distribution.

From Sigmoid to Logit

- Assume x is input
  - Called logit here
- We applied sigmoid(x) to get y
- Now, inverse the y function to get x from y
  - Think a = sqrt(b)
  - b = a^2
- Apply the log to cancel exp

```
y = 1/(1 + exp(-x))

1 + exp(-x) = 1/y

exp(-x) = 1/y - 1

exp(-x) = 1/y - y/y

exp(-x) = (1 - y)/y

ln(exp(-x)) = ln((1 - y)/y)

-x = ln((1 - y)/y)

x = -ln((1 - y)/y)

x = ln(y/(1 - y))
```

src

From Logit to Sigmoid

- Apply the exp to cancel log and extract x (pi)

$$logit(\pi) = t = \log \frac{\pi}{1-\pi}$$

$$e^t = \frac{\pi}{1-\pi} \Rightarrow (1-\pi) \cdot e^t = \pi$$

$$\Rightarrow (1+e^t) \cdot \pi = e^t$$

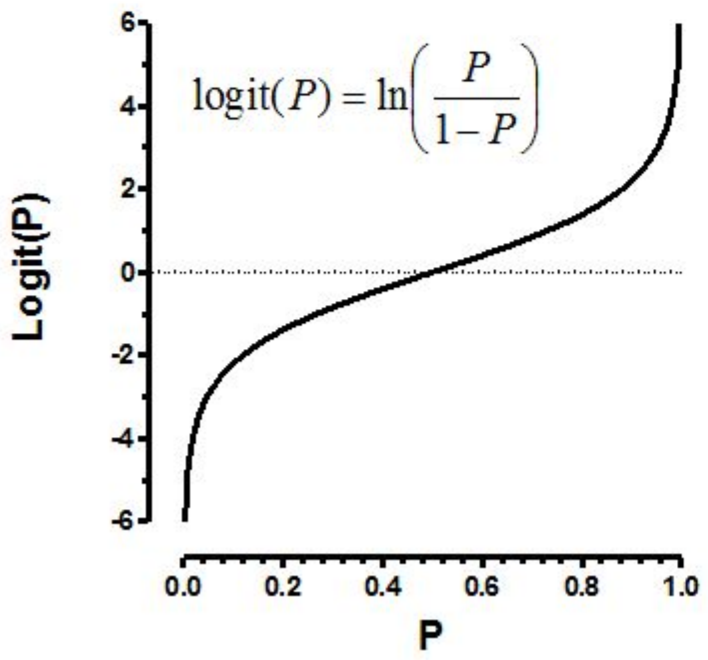
$$\Rightarrow \pi = \frac{e^t}{(1+e^t)} = \frac{1}{(1+e^{-t})} = \sigma(t)$$

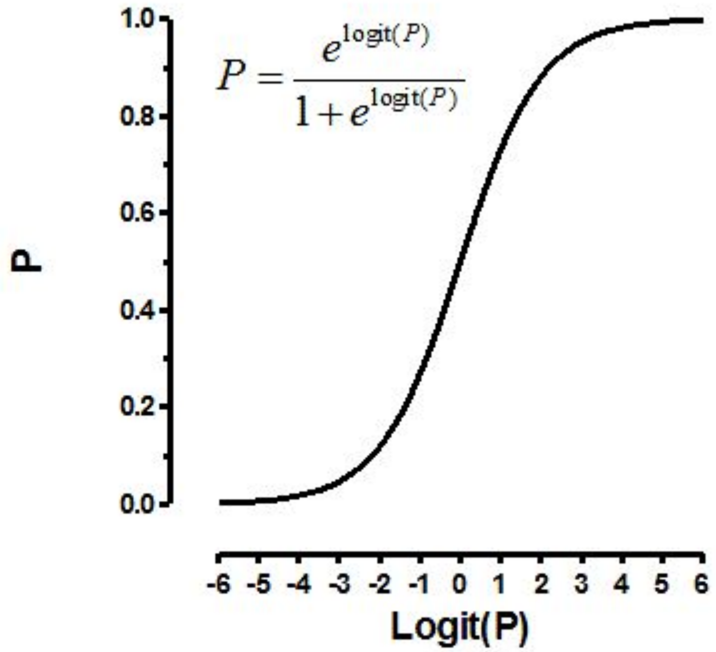
src

Probability, log-odds, and odds

- Assume the probability of an event is $p = 0.2$
- odds = $p / (1-p) = 0.2 / 0.8 = 0.25$
- Logit = log-odds = $\ln(0.25) = -1.3863$
- Probability from odds = odds / (1+odds) = $0.2/1.25 = 0.2$
- Probability from log-odds:
  $\exp(\ln(odds)) / (1 + \exp(\ln(odds))) =$
  $\exp(-1.3683) / (1 + \exp(-1.3683)) = 0.2$

src





- For probability 0.5 ⇒ logit = ln(0.5/0.5) = ln(1) = 0
    - Remember this for next lecture
- For probability 0.7⇒ logit = ln(0.7/0.3) = 0.85
- For probability 0.9⇒ logit = ln(0.9/0.1) = 2.2
- Both curves are increasing functions

src

“Acquire knowledge and impart it to the people.”

“Seek knowledge from the Cradle to the Grave.”