CS 145 Discussion 3

Reminders

- HW2 will be released today (10/20/2017)
  - HW2 out, due 10/29/2017 11:59 pm (Sunday)
  - It would be easier and clearer than HW1.

- Data crawler for the course project
  - Start implementing the crawler as soon as possible
  - Need time to crawl sufficient data

Today's Outline

- Support Vector Machine
  - Recap
  - Example
- Neural Network
  - Backpropagation derivation
  - Exploding and vanishing gradients
- NN Examples
  - Binary classification
  - Multi-class classification
  - Multi-label classification

Support Vector Machine

Support Vector Machine Recap

- Hyperplane separating the data points

$$\boldsymbol{w}^{\boldsymbol{T}}\mathbf{x} + \mathrm{b} = 0$$

- Maximize margin

$$\rho = \frac{2}{\|w\|}$$

- Solution

$$\boldsymbol{w} = \sum \alpha_i y_i \mathbf{x}_i \qquad b = \sum_{k:\alpha_k \neq 0} (y_k - \boldsymbol{w}^{\boldsymbol{T}}\mathbf{x}_k)/N_k$$

Margin Formula

- Margin Lines

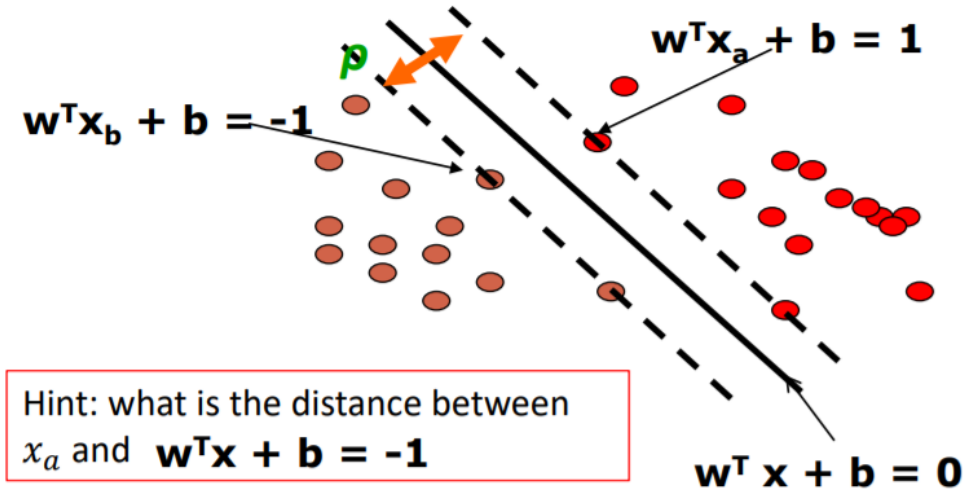
$$\boldsymbol{w^T x_a} + \mathrm{b} = 1 \qquad \boldsymbol{w^T x_b} + \mathrm{b} = -1$$

- Distance between parallel lines

$$d = \frac{|c_2 - c_1|}{\sqrt{a^2 + b^2}}$$

- Margin

$$\rho = \frac{|(b+1) - (b-1)|}{\|w\|} = \frac{2}{\|w\|}$$



Linear SVM Example

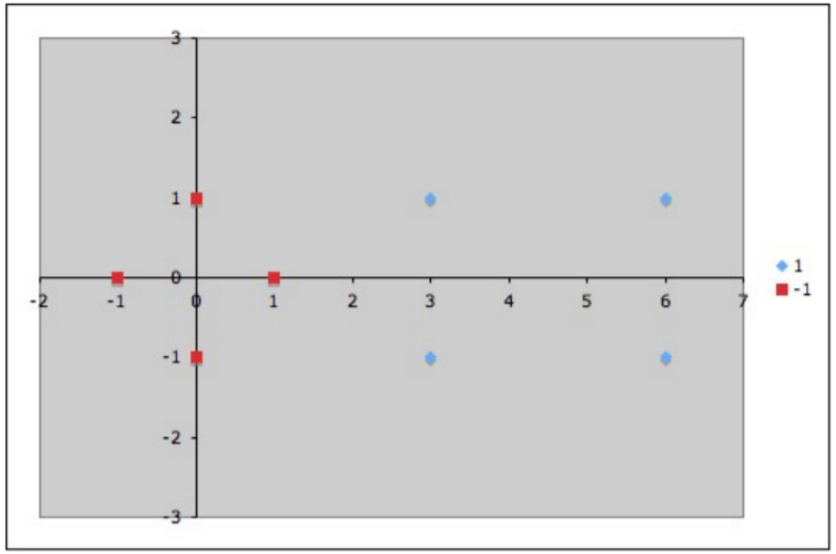
- Positively labeled data points (1 to 4)

$$\left\{\begin{pmatrix}3\\1\end{pmatrix},\begin{pmatrix}3\\-1\end{pmatrix},\begin{pmatrix}6\\1\end{pmatrix},\begin{pmatrix}6\\-1\end{pmatrix}\right\}$$

- Negatively labeled data points (5 to 8)

$$\left\{\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}-1\\0\end{pmatrix}\right\}$$

- Alpha values
  - $\alpha_1 = 0.75$
  - $\alpha_2 = 0.75$
  - $\alpha_5 = 3.5$
  - Others = 0



Linear SVM Example

- Which points are support vectors?
- Calculate normal vector of hyperplane: $\boldsymbol{w}$
- Calculate the bias term
- What is the decision boundary?
- Predict class of new point (4, 1)

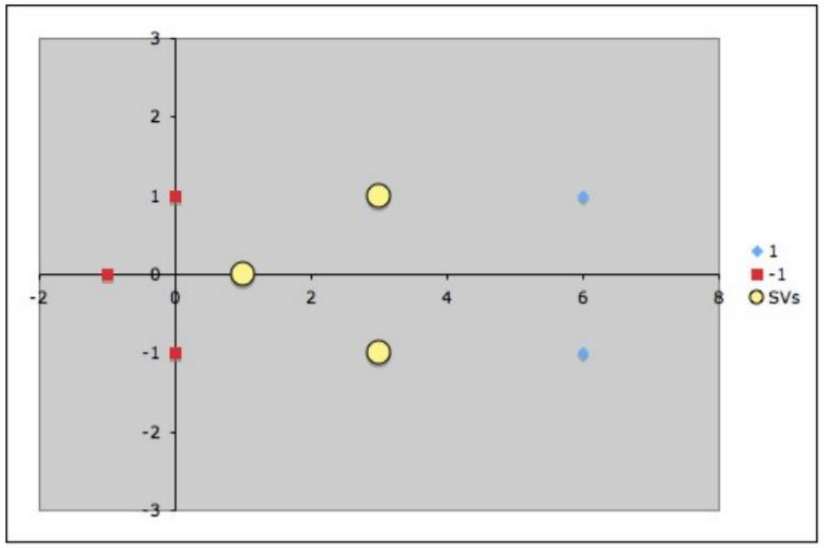
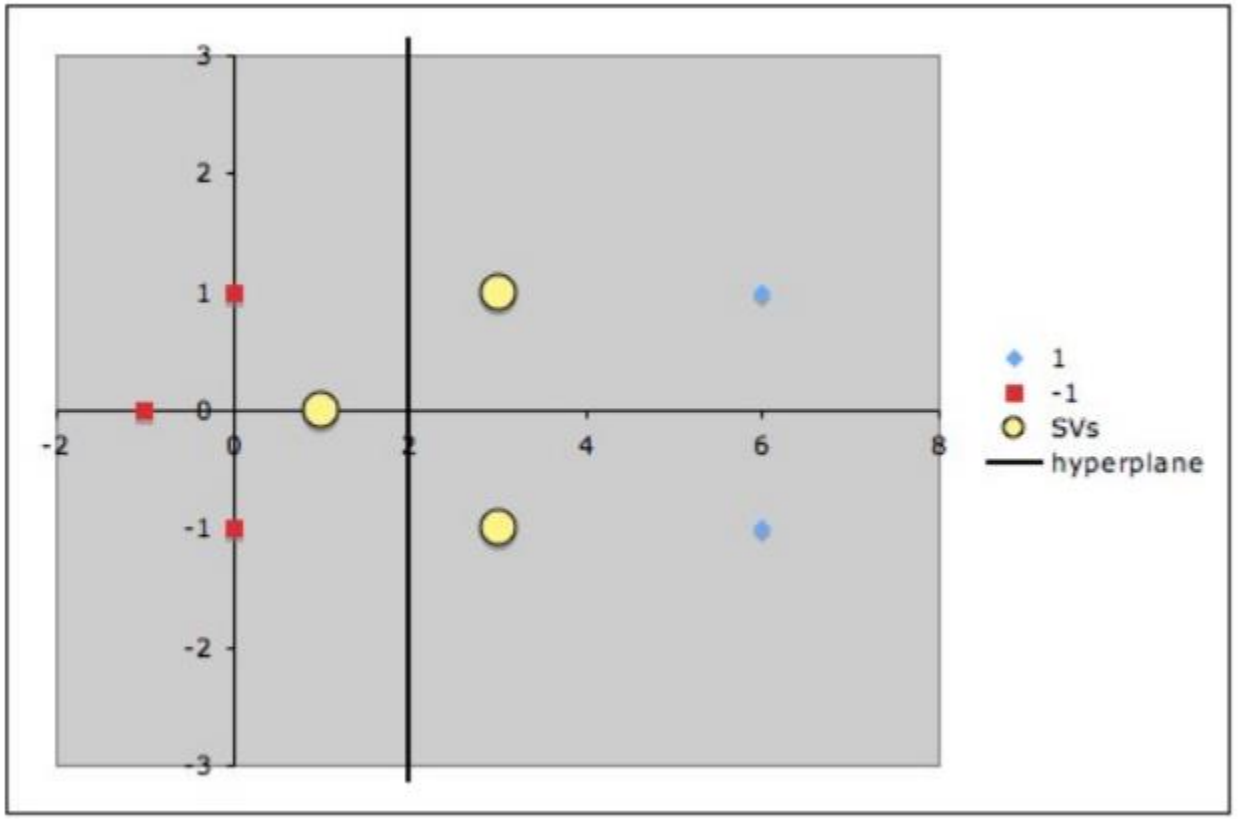


Figure 2: The three support vectors are marked as yellow circles.

Plot



Non-linear SVM Example

- Positively labeled data points (1 to 4)

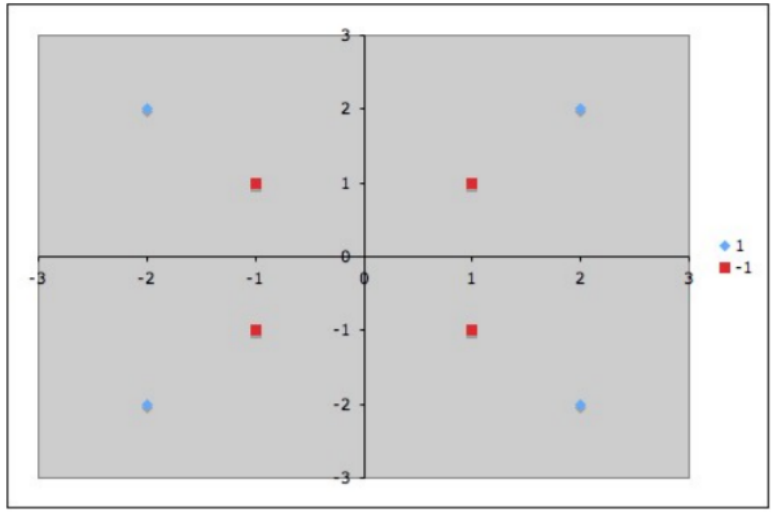
$$\left\{\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}2\\-2\end{pmatrix},\begin{pmatrix}-2\\-2\end{pmatrix},\begin{pmatrix}-2\\2\end{pmatrix}\right\}$$

- Negatively labeled data points (5 to 8)

$$\left\{\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}-1\\1\end{pmatrix}\right\}$$

- Non-linear mapping

$$\Phi_1\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{cases}\begin{pmatrix}4-x_2\\4-x_1\end{pmatrix} & \text{if } \sqrt{x_1^2+x_2^2}>2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix} & \text{otherwise}\end{cases}$$



Non-linear SVM Example

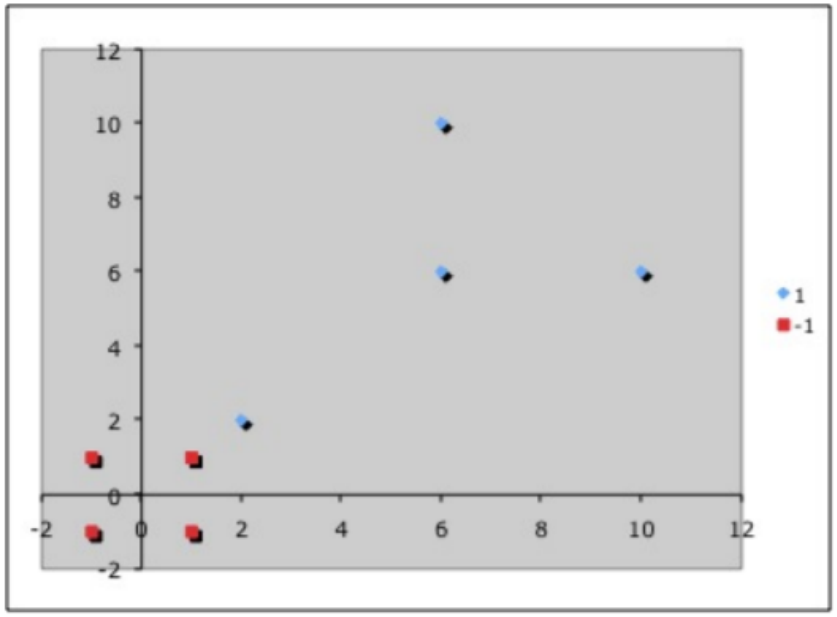
- New positively labeled data points (1 to 4)

$$\left\{\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}6\\2\end{pmatrix},\begin{pmatrix}6\\6\end{pmatrix},\begin{pmatrix}2\\6\end{pmatrix}\right\}$$

- New negatively labeled data points (5 to 8)

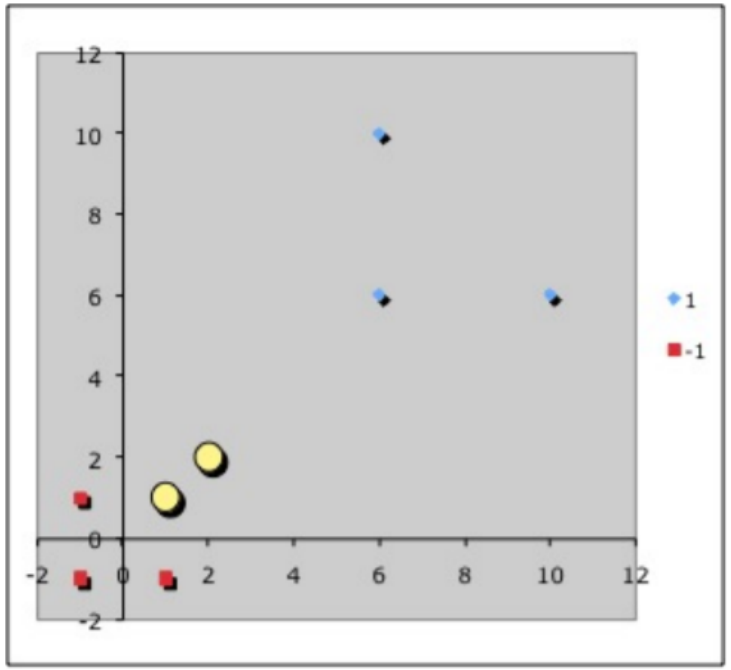
$$\left\{\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}-1\\1\end{pmatrix}\right\}$$

- Alpha values
  - $\alpha_1 = 4$
  - $\alpha_5 = 7$
  - Others = 0

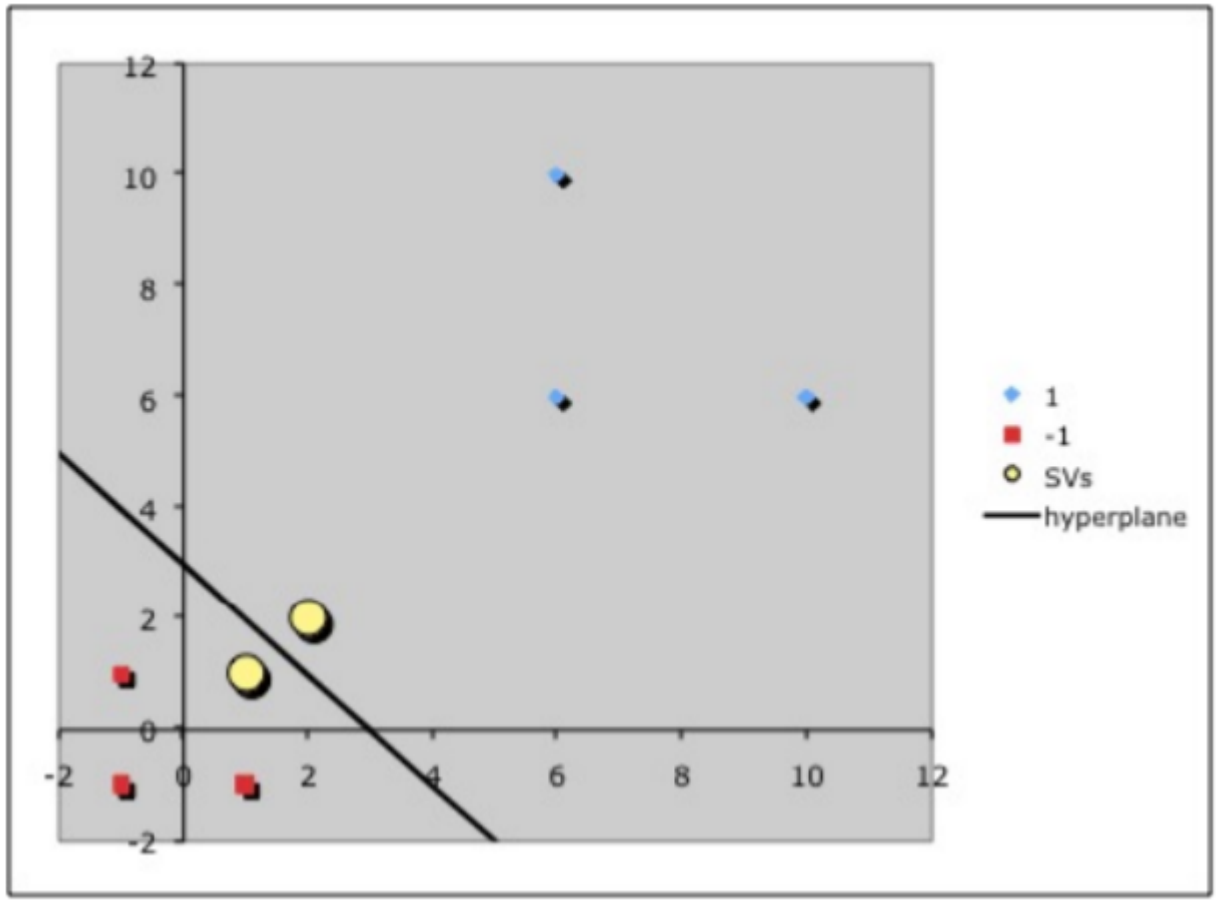


Non-linear SVM Example

- Which points are support vectors?
- Calculate normal vector of hyperplane: $\boldsymbol{w}$
- Calculate the bias term
- What is the decision boundary?
- Predict class of new point (4, 5)



Plot



Backpropagation in Neural Network

Backpropagation Derivation

- Equations

  Backpropagate the error (by updating weights and biases)

  - For unit j in output layer: $Err_j = O_j(1 - O_j)(T_j - O_j)$
  - For unit j in a hidden layer: : $Err_j = O_j(1 - O_j)\sum_k Err_k w_{jk}$
  - Update weights: $w_{ij} = w_{ij} + \eta Err_j O_i$
  - Update bias: $\theta_j = \theta_j + \eta Err_j$

- Derivation (pdf also uploaded on CCLE)
  - https://www.cs.swarthmore.edu/~meeden/cs81/s10/BackPropDeriv.pdf

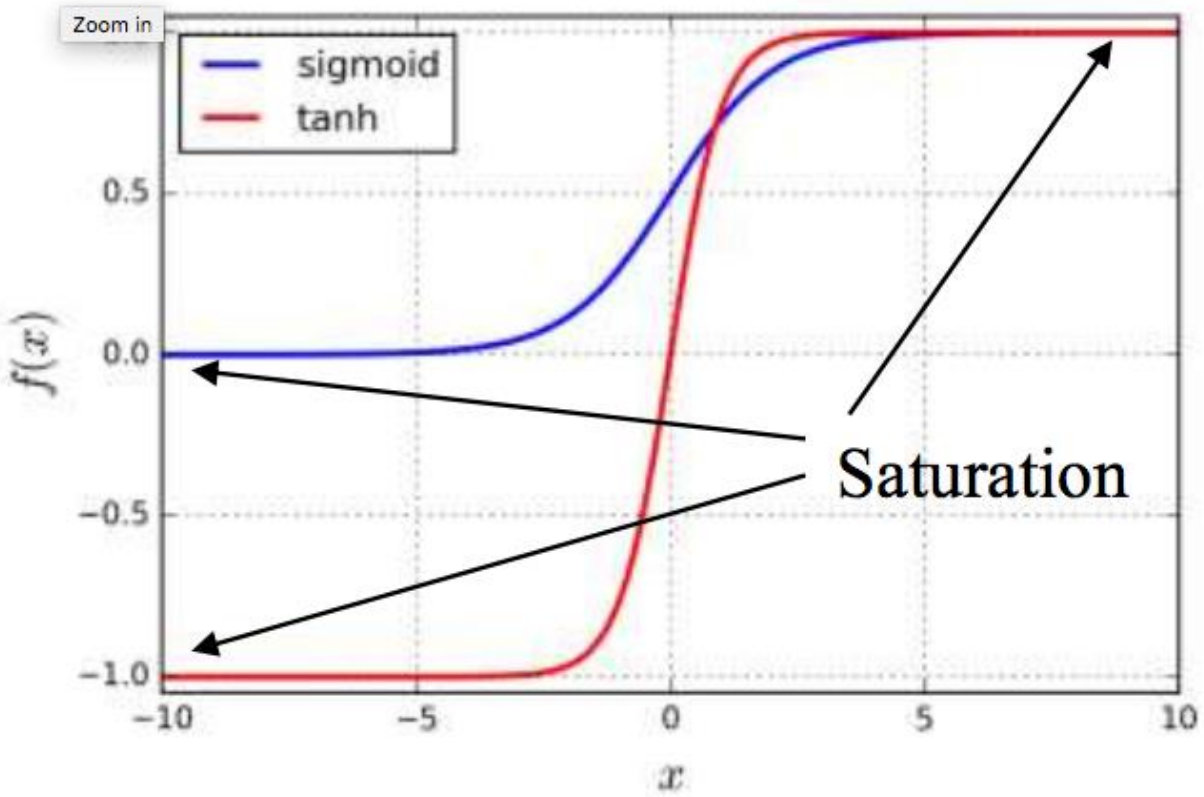
Exploding and vanishing gradients in Neural Network

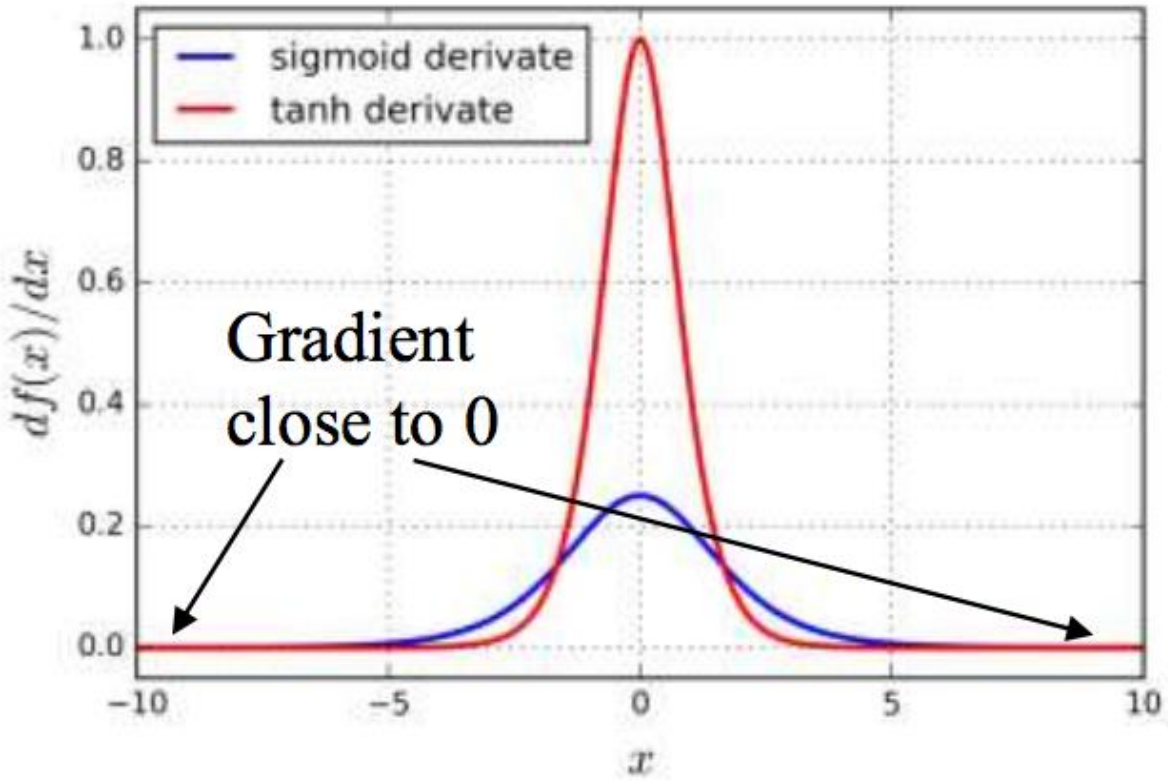
Vanishing Gradient Problem

- Deep neural networks use backpropagation a lot.
- Backpropagation applies chain rule
- The chain rule multiplies derivatives.
- Often these derivatives between 0 and 1.
- As the chain gets longer, products get smaller until they disappear.

$$\frac{dJ}{d\theta^{(l)}} = \frac{dJ}{da^{(L)}} \cdot \frac{da^{(L)}}{dz^{(L)}} \cdot \frac{dz^{(L)}}{da^{(L-1)}} \cdot \frac{da^{(L-1)}}{dz^{(L-1)}} \cdots \frac{da^{(l+1)}}{dz^{(l+1)}} \cdot \frac{dz^{(l+1)}}{d\theta^{(l)}}$$

Vanishing Gradient Problem (Cont'd)

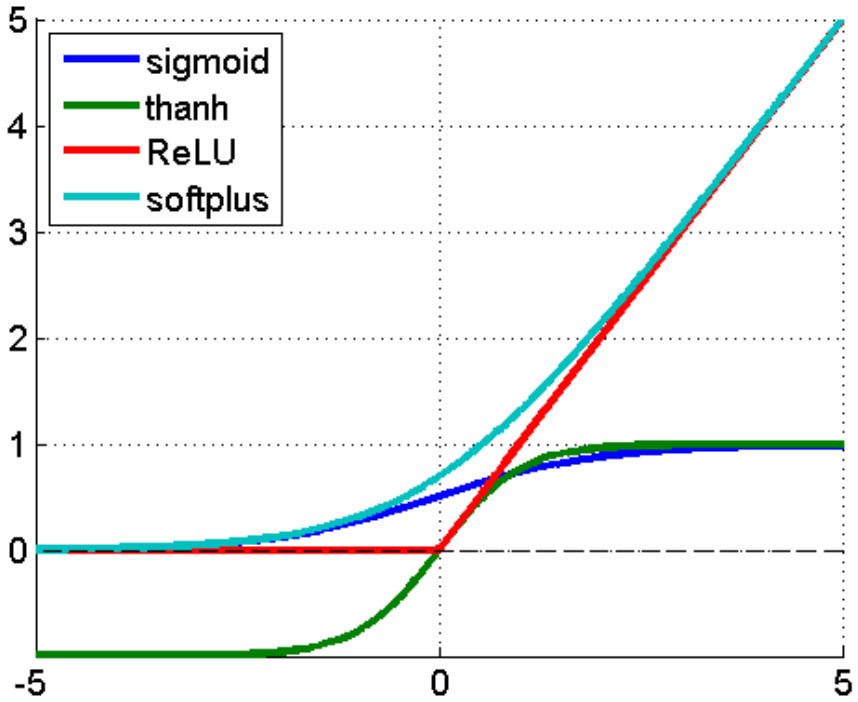




Solution: Other Activation Function

- RELU(Rectified Linear Unite)
- $f_{ReLU}(x) = \max(0, x)$

- Does not vanish as x increases
- Faster without computing exponential functions



Exploding Gradient Problem

- With gradients larger than 1, products may become larger and larger as the chain becomes longer and longer
- Causing overlarge updates to parameters

- Solution:
  - Gradient clipping (limiting the gradients)
  - Reduce learning rate
  - Add regularization as constraints on weights

Solve Classification Problems with Neural Networks

Classification Tasks

- Binary classification
  - Single output with two possible values
  - E.g., Yes or No, Rain or Not
- Multi-class classification
  - Single outputs with multiple possible values
  - E.g., Rain/Sunny/Cloudy
- Multi-label classification
  - Multiple outputs, each of them is a binary
  - E.g., genres prediction

The Final Layer of Neural Network

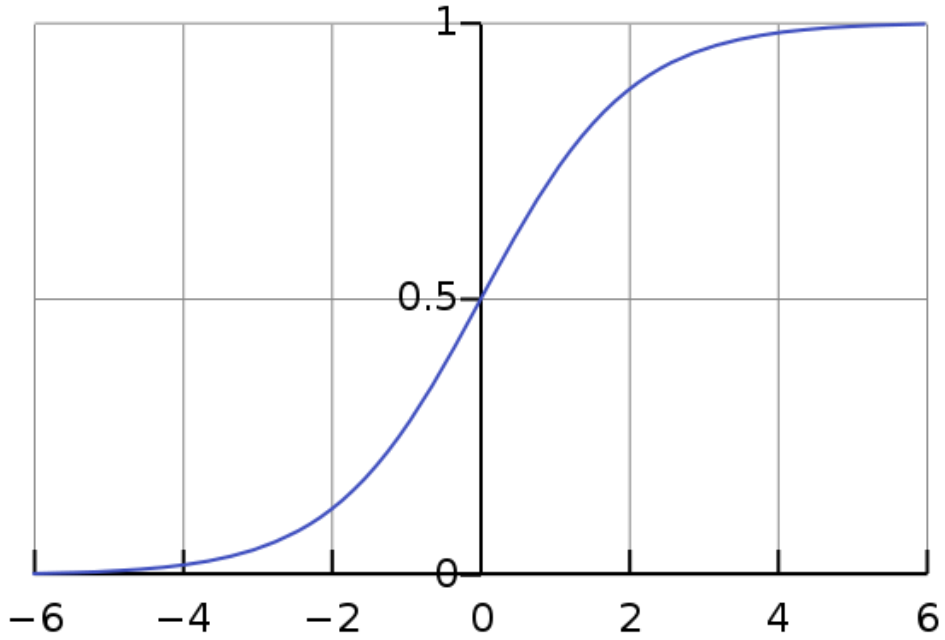
Binary Classification

- Estimate the probability of belonging a certain class

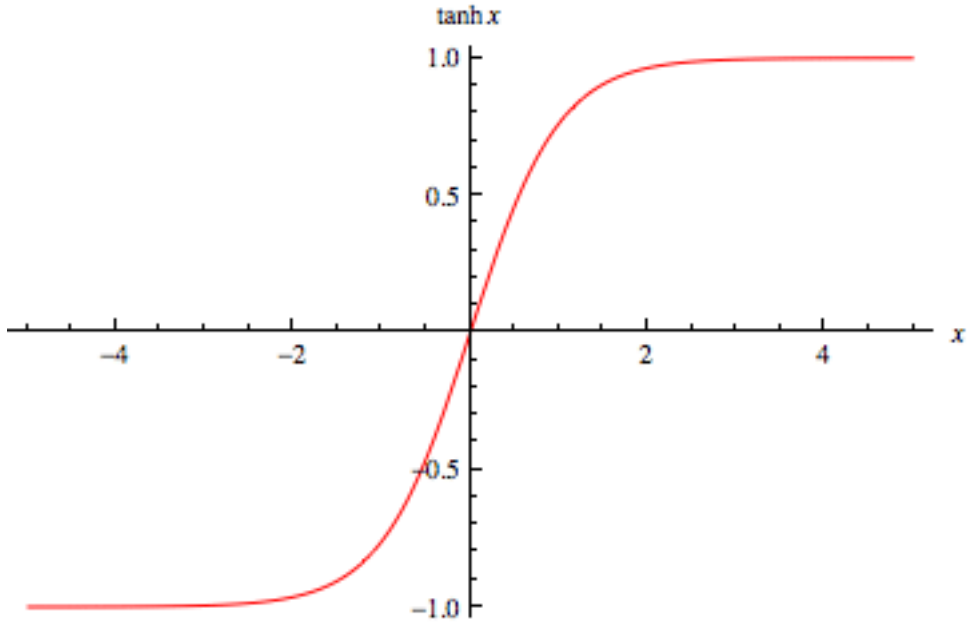
$$P(Y = 0 \mid X, W)$$

- Sigmoid function
  - Logistic $\frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$
  - Hyperbolic tangent (tanh) $\frac{e^{2x}-1}{e^{2x}+1}$

Sigmoid Functions

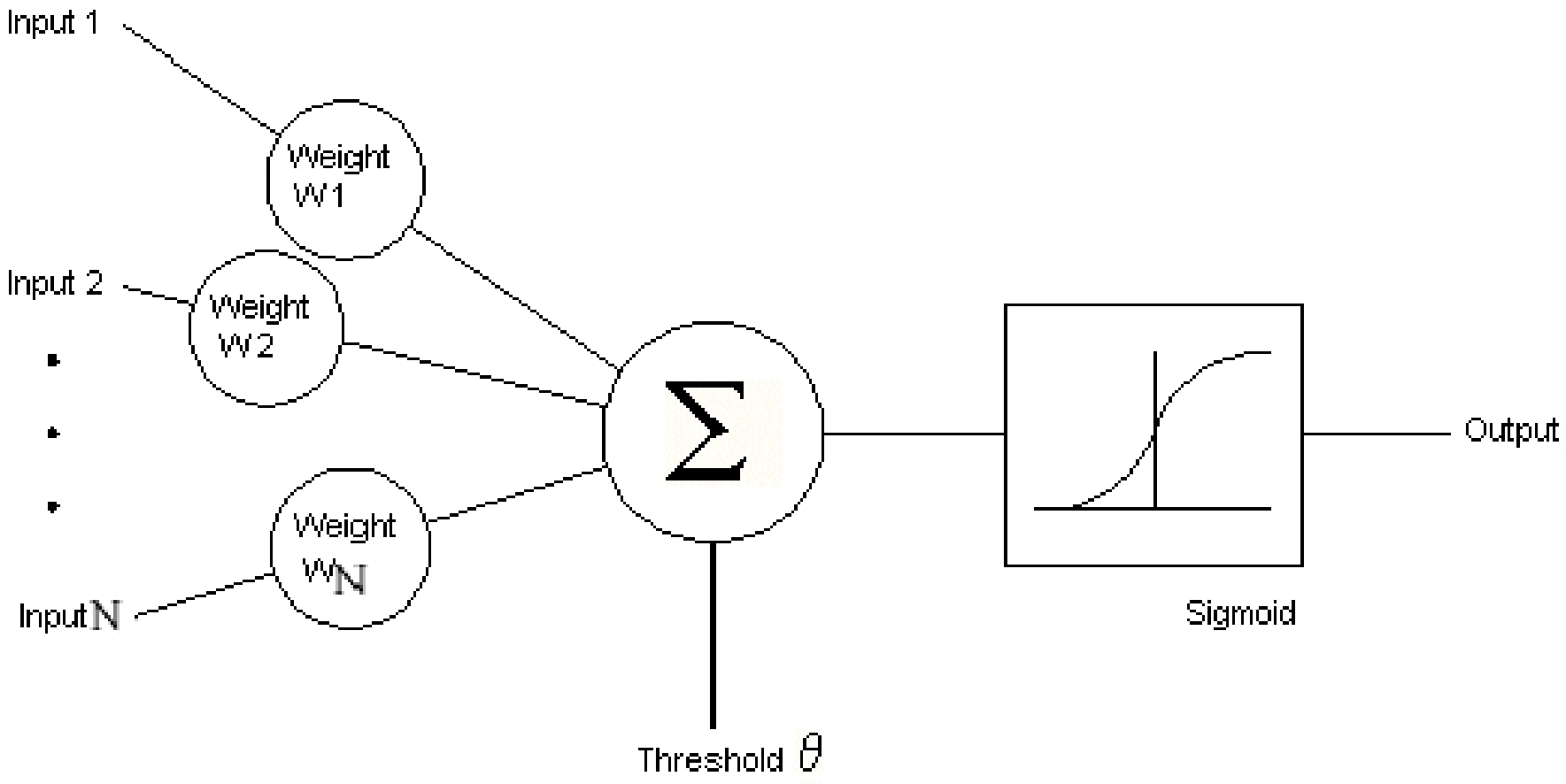


Logistic



Hyperbolic tangent (tanh)

Binary Classification (Cont'd)

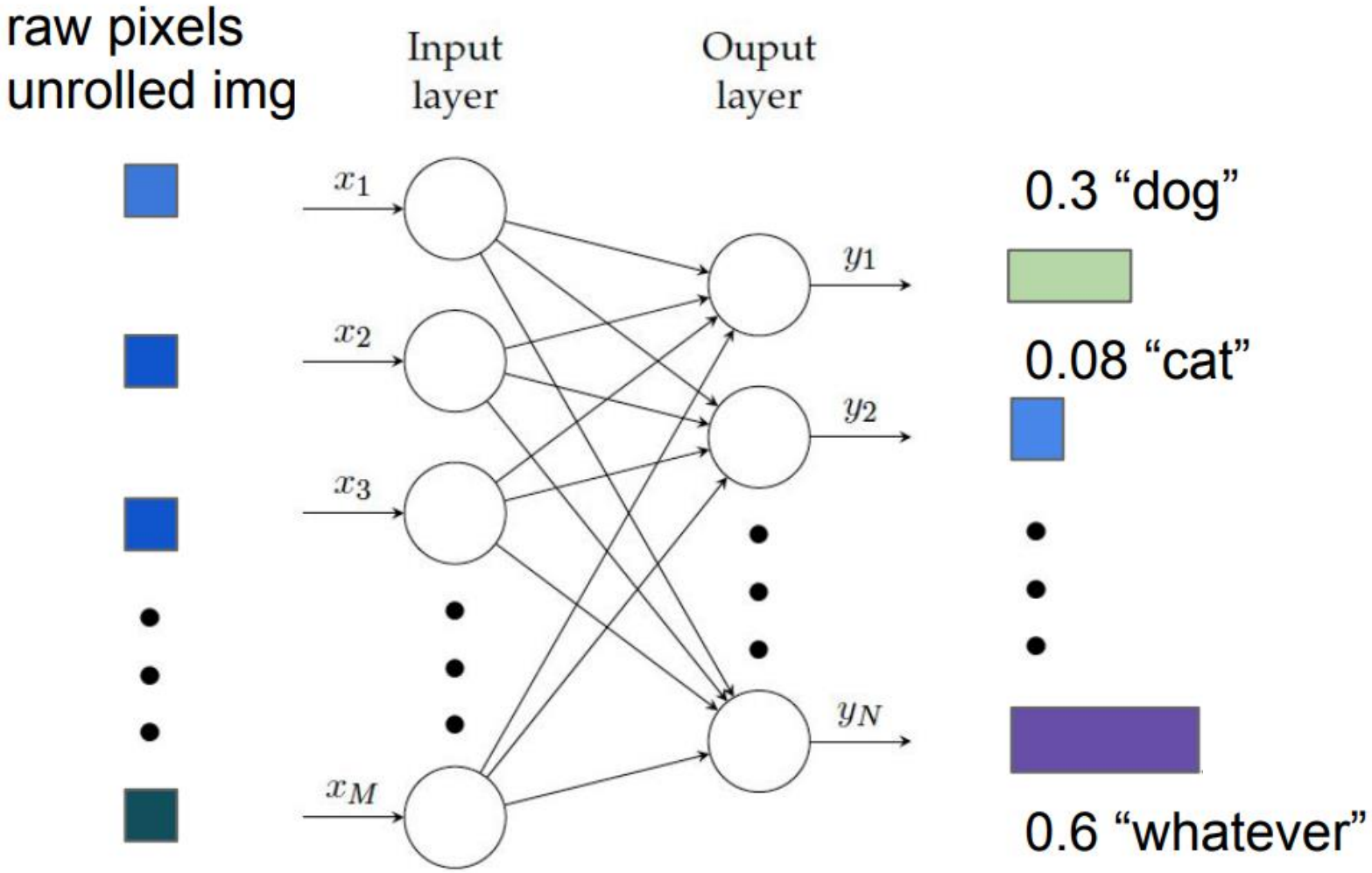


Multi-class Classification

- Predict the probabilistic distribution of classes
- Softmax Function or Normalized Exponential Function
  - Generate a score based on the exponential function for each class
  - Normalize scores as the probabilistic distribution

$$P(y = j \mid x) = \frac{e^{x^T w_j}}{\sum_k e^{x^T w_k}}$$

An example of image classification



Multi-label Classification

- Output independent probabilities
  - K sigmoid outputs
- Share weights for different binary prediction