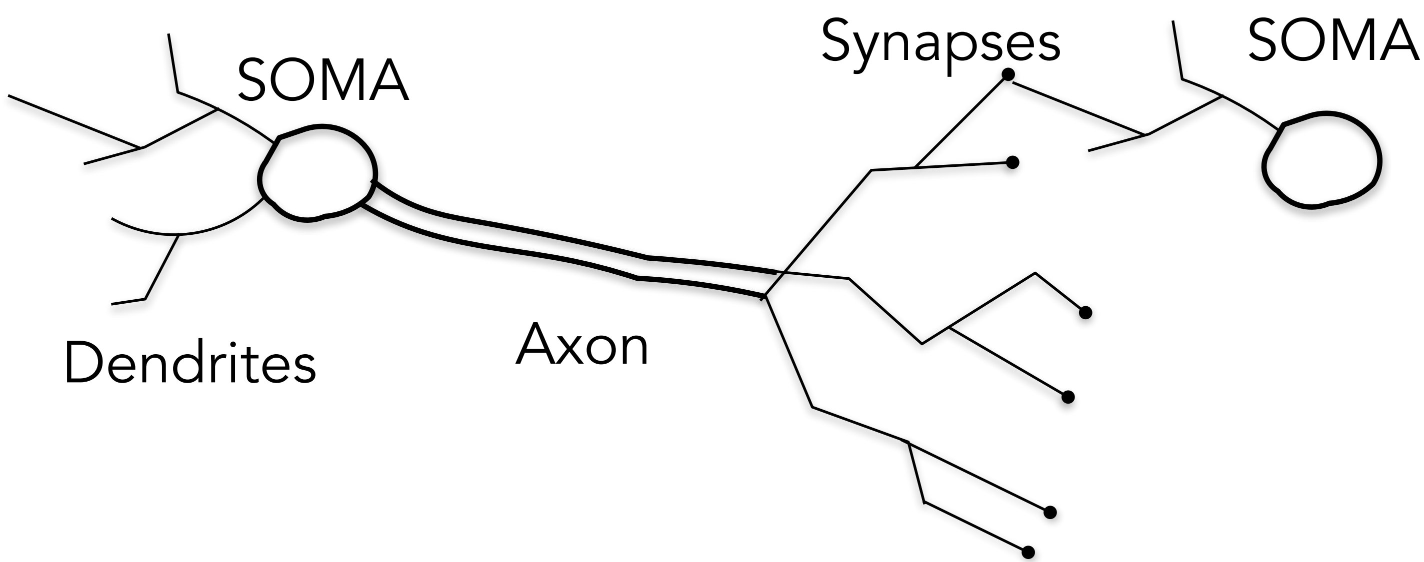
Lecture 15-16
Autoencoders

Ref: Outlier Analysis, Charu C Agrawal
Ref: Bishop, Christopher M. Pattern recognition and machine learning.
Ref: Tutorial -https://web.mit.edu/zoya/www/SVM.pdf

Artificial Neural Networks

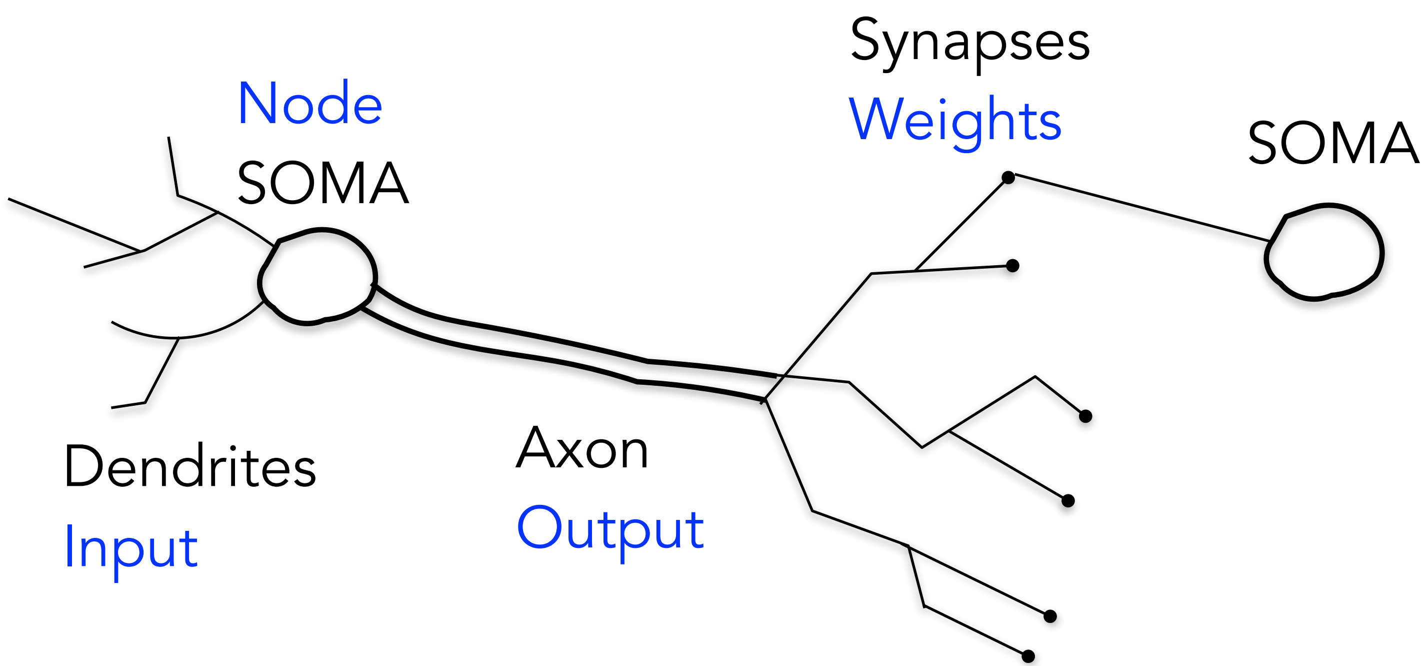
Every encoding-based anomaly detection method assumes that the data has redundancy , hence, we can represent the data in lower dimension.

Biological Neurons



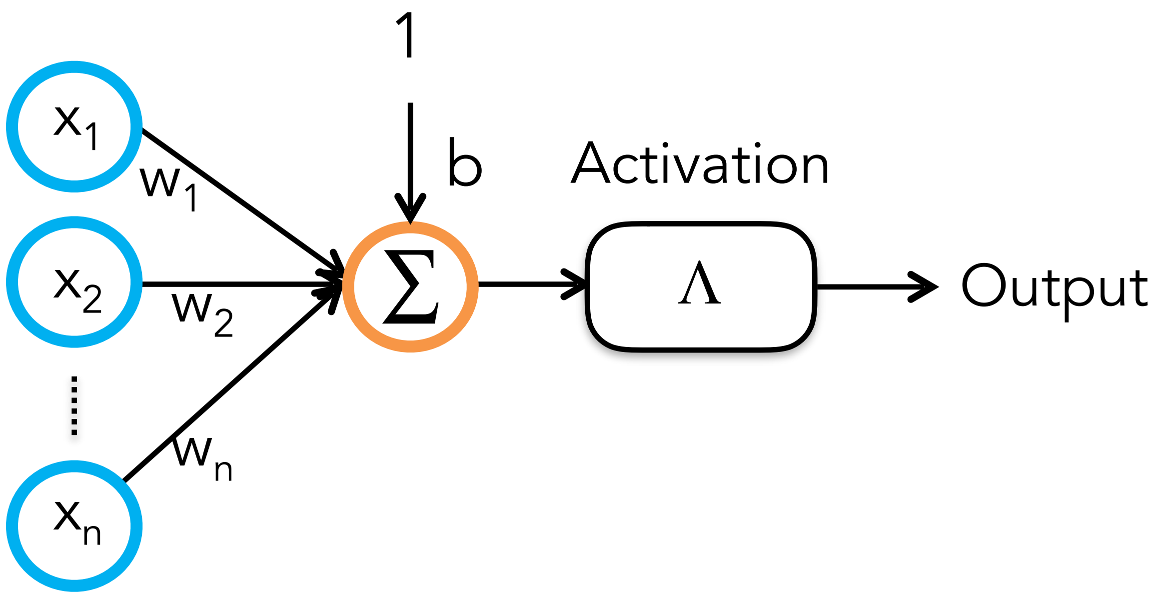
- **Information processing unit of the brain**
- **Connected by Synapses**

Artificial Neurons



- **Information processing unit of the brain**
- **Connected by Synapses**

Artificial Neural Network



$$y_{in} = x_1 w_1 + x_2 w_2 + \ldots x_n w_n + b$$

$$Y = \Lambda(y_{in})$$

$$E.g. \quad Y = \frac{1}{1 + e^{-y_{in}}}$$

Activation Function

- An additional effort to get exact output
- Makes the system non-linear
- E.g. binary sigmoidal function of bipolar sigmoidal function
- Newer activation functions: ReLu, Tanh, softmax

Unipolar Vs Bipolar Activation Function

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

$$f'(x) = 2 * f(x) - 1 = \frac{2}{1 + e^{-x}} - 1 = \frac{1 - e^{-x}}{1 + e^{-x}}$$

Bias

- Without bias, the model will train over points passing through origin only
- Bias units are not connected to the previous layers
- It is represented by a weight of a node whose value is always 1

Weights

- Contain actual knowledge of the data
- They define the steepness of the activation function
- In other words, they determine how fast the activation function will trigger with change in the input

Learning Parameters (Weights)

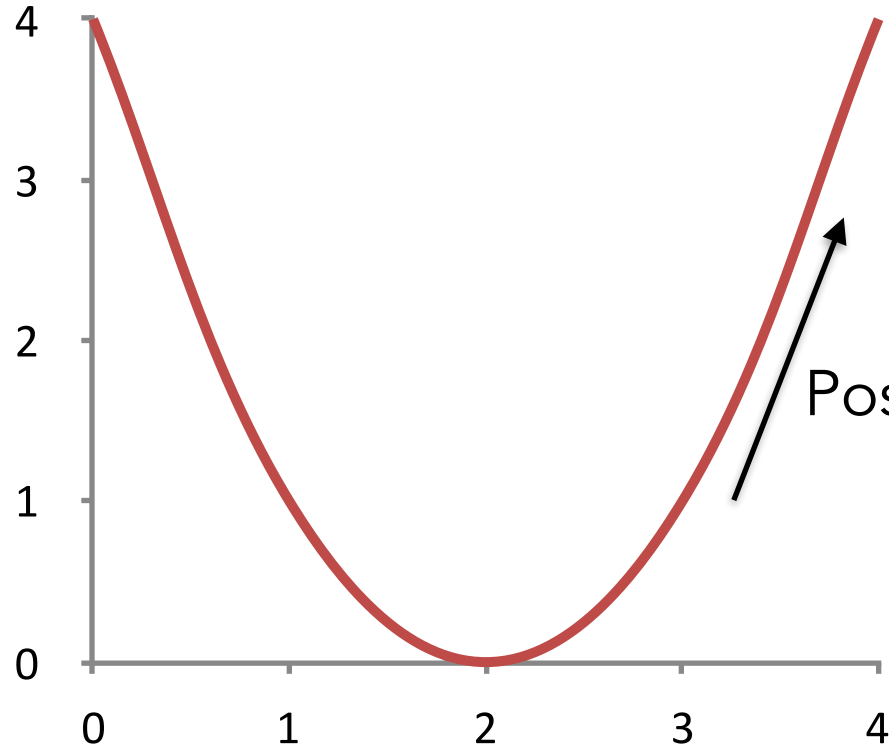
Gradient descent & back-propagation

- Use gradient descent to calculate the below functions roots

$$f(w) = (w - 2)^2$$

$$w_t = w_{t-1} - \eta \frac{\partial f(w)}{\partial w}$$

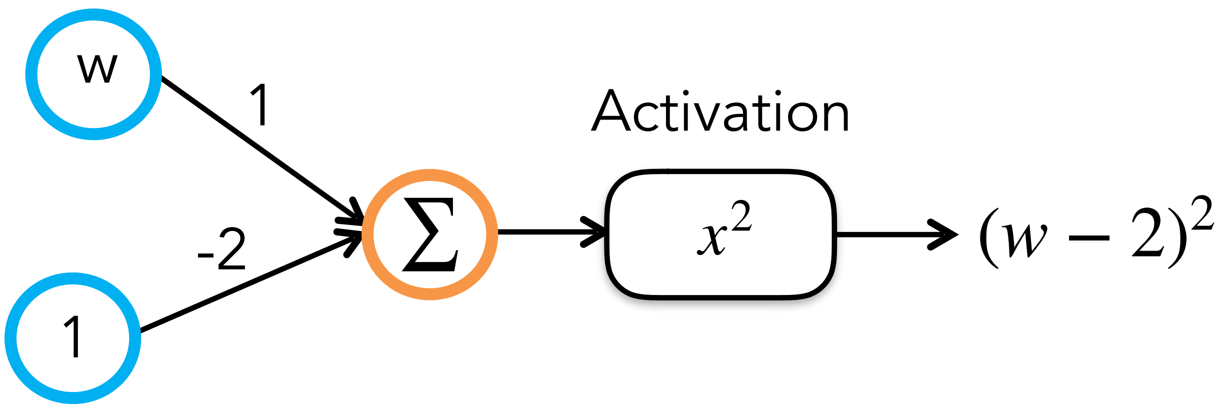
Gradient descent & back-propagation



Positive $\frac{\partial f(w)}{\partial w}$

Can you draw the equivalent neural network?

$$f(w) = (w-2)^2$$



Neural networks represent a function, whole training process is actually function approximation!

ANN Training

- Need labeled training set (expected output with each input)
- Start with random initial weights
- Use the gradient descent to adjust the weights
- Do for all samples
- Repeat multiple times for the whole dataset - multiple epochs

Gradient Descent Variants

- Vanila: calculate error over all samples and then update weights
- Stochastic Gradient Descent: update weights for each sample
- Batch gradient descent: calculate error over a n samples and then update weights

The whole neural network represents a function with single/multiple input and single/multiple outputs.

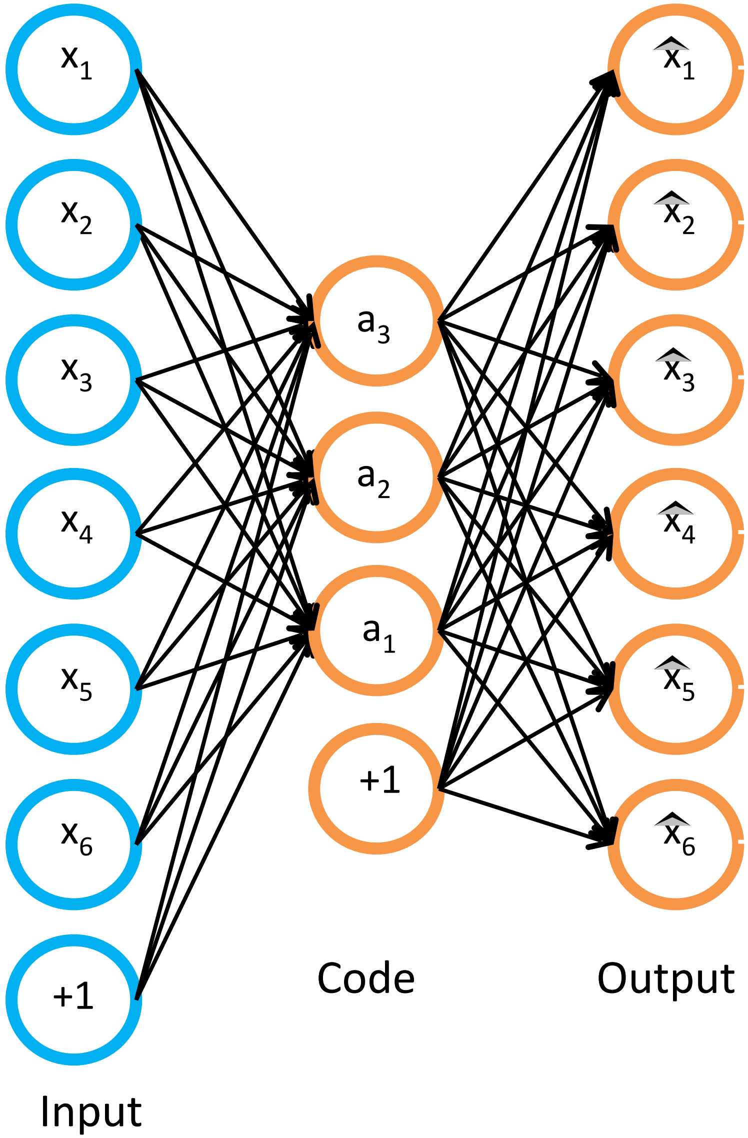
Main Idea

- Train the encoder to compress the normal data.
- When used on normal data, we will get low reconstruction error.
- When used on the abnormal data, the reconstruction error will be large.

Autoencoder Networks

- The aim of an auto encoder is to learn a representation of the given set of data.
- Mainly for the purpose of dimensionality reduction.

Autoencoder Networks



Number of Layers

- Always has odd number of layers in total
- The middle layer represents the code
- A deeper architecture can learn more complex data
- Layers may have non-linear activation function
- Number of layers should be the same in encoder and decoder, but the activation functions can be different

Activation Function

- Sigmoid functions are undoubtedly the most common activations in AEs.
- The activation function brings the nonlinearity to the compression
- Without the activation function, the auto encoder is equivalent to PCA with Eigen directions equal to the number of nodes in the middle layer

Non-differentiable Activation Function

Non-differentiable activation functions, such as ReLU, are generally not preferred in autoencoder as they make the reconstruction difficult

Loss Function/Objective Function

- A typical objective function is MSE
- Another possibility is cross entropy loss

Training

- The most common training approach for auto encoders is stochastic gradient descent
- Sometimes regularisation term (forcing smaller weights) is introduced to avoid overfitting
- AEs can also be trained layer-by-layer

Anomaly Score

- The AE is trained to have small reconstruction error over normal data
- In auto-encoders, anomaly score is the reconstruction error/loss function value/ objective function value