CS 6820 – Machine Learning

Lecture 3

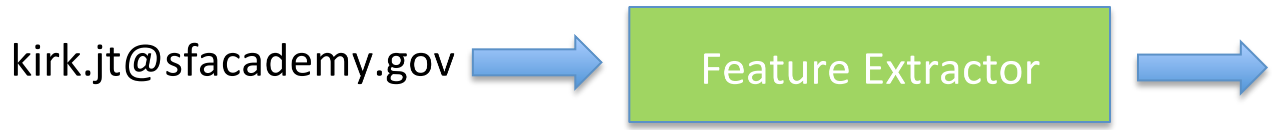
Instructor: Eric S. Gayles, PhD.

Jan 10, 2018

Feature Extraction

- Example task: Determine whether a string ***x*** is an email address.
- Prediction = ***y***.
- Question: What properties of ***x*** might be relevant for predicting y?
- Feature extractor: Given input ***x***, output a set of (feature_name, feature_value) pairs.

Feature Extraction



length.gt.10	1
fracOfAlpha	100
containsAtSign	1
endsWithDotCom	0
endsWithDotOrg	0
endsWithDotGov	1

$$\begin{bmatrix} 1 \\ 100 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

For an input x, its feature vector is:

$$\phi(x) = [\phi_1(x), \ldots, \phi_d(x)].$$

Think of $\phi(x) \in \mathbb{R}^d$ as a point in a high-dimensional space.

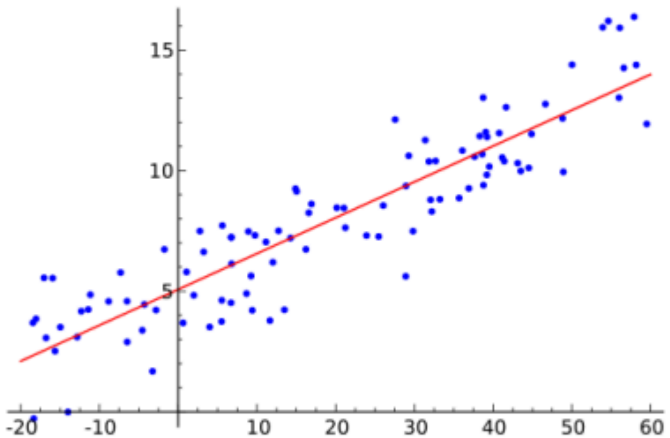
Loss Function Definition

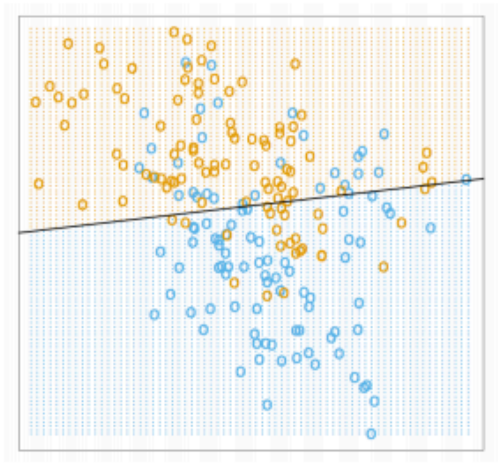
- A loss function Loss(x, y, w) quantifies how unhappy you would be if you used w to make a prediction on x when the correct output is y. It is the object we want to minimize.

* Liang & Ermon

Linear Regression – Dual Roles

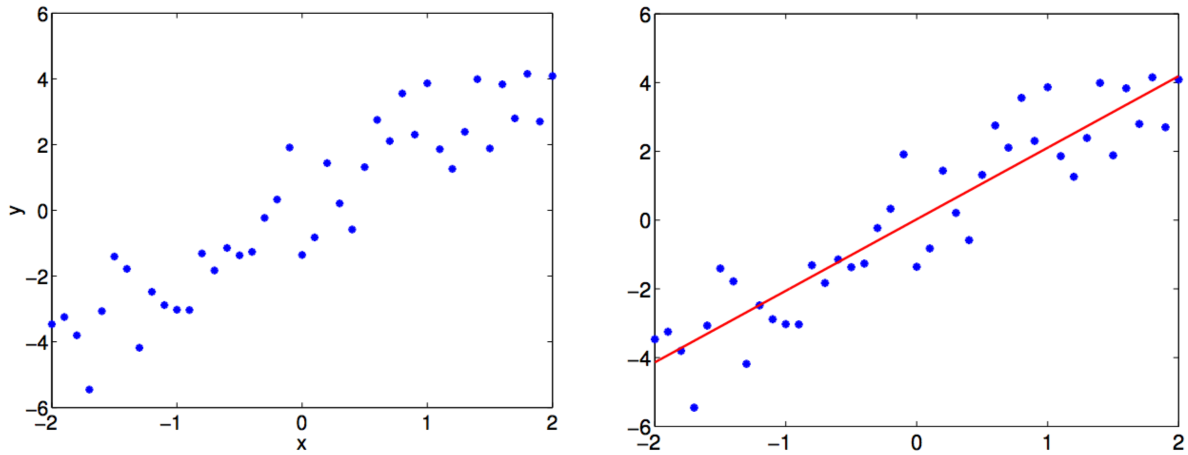
- regression problem ⇝ variable to predict is continuous/quantitative
- classification problem ⇝ variable to predict is discrete/qualitative





* Chiarandini

Linear Regression in the ML Context



- We need to define a class of functions (types of predictions we will try to make) such as linear predictions

$$f(x; w_1, w_0) = w_0 + w_1 x$$

where w_1, w_0 are the *parameters* we need to set.

Linear Regression

- An equation can be fit to show the best linear relationship between two variables:

$$\boxed{Y = \beta_0 + \beta_1 X}$$

Where Y is the dependent variable and
X is the independent variable
β_0 is the Y-intercept
β_1 is the slope

* Basic Business Statistic, Berenson et. al.

Linear Regression

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X
- Linear regression population equation model

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

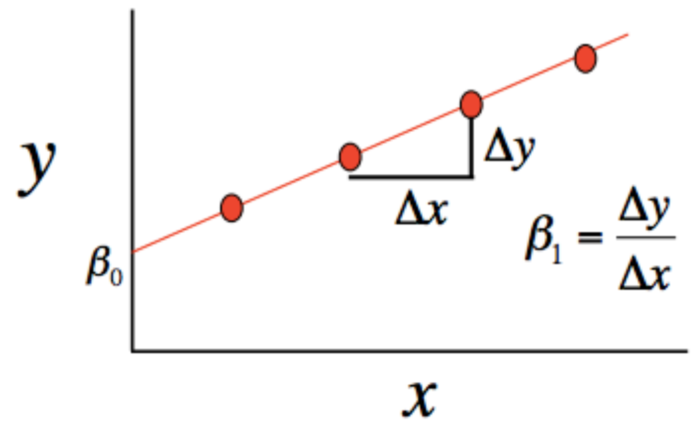
* Basic Business Statistic, Berenson et. al.

Linear Regression in the ML Context

- Much of mathematics is devoted to studying variables that are deterministically related to one another

* Basic Business Statistic, Berenson et. al.

$$y = \beta_0 + \beta_1 x$$



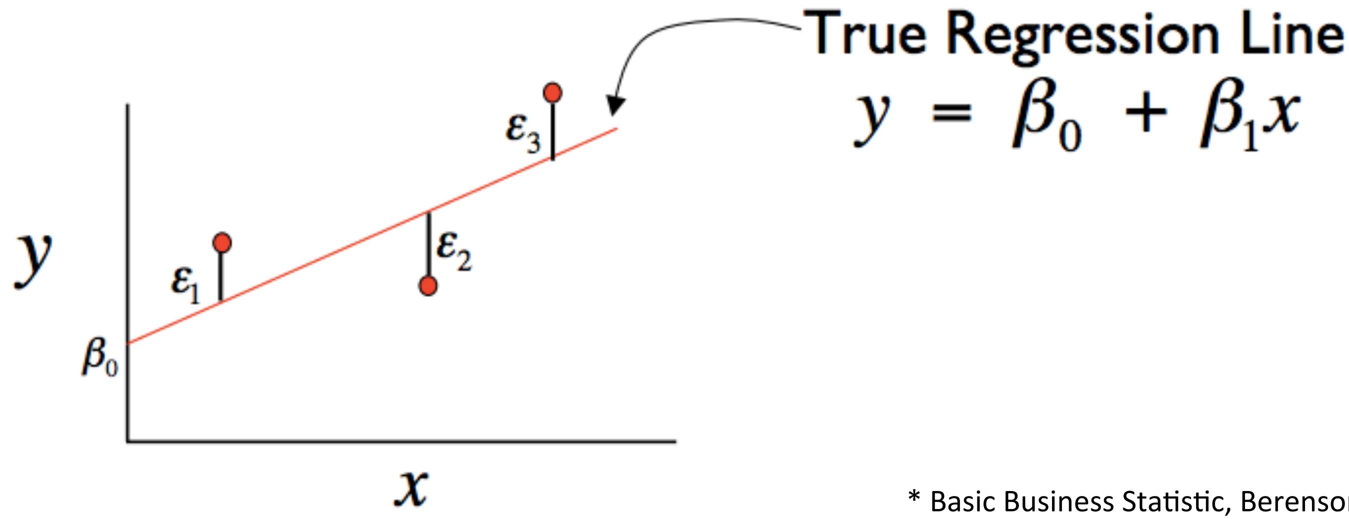
- But we're interested in understanding the relationship between variables related in a nondeterministic fashion

Linear Regression in the ML Context

- Definition: There exists parameters β_0, β_1, and σ^2, such that for any fixed value of the independent variable x, the dependent variable is related to x through the model equation

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- ε is a rv assumed to be N(0, σ^2)



* Basic Business Statistic, Berenson et. al.

Linear Regression in the ML Context

- ***Predicted***, or fitted, values are values of y predicted by the least-squares regression line obtained by plugging in $x_1, x_2, \ldots, x_n$ into the estimated regression line

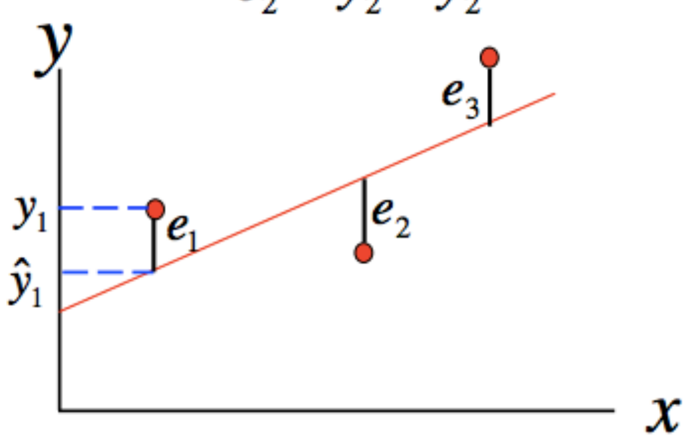
$$\hat{y}_1 = \hat{\beta}_0 - \hat{\beta}_1 x_1$$

$$\hat{y}_2 = \hat{\beta}_0 - \hat{\beta}_1 x_2$$

- ***Residuals*** are the deviations of observed and predicted values

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$



* Basic Business Statistic, Berenson et. al.

Linear Regression in the ML Context

- They allow us to calculate the error sum of squares (SSE):

$$SSE = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Which in turn allows us to estimate σ^2:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

- As well as an important statistic referred to as the coefficient of determination:

$$r^2 = 1 - \frac{SSE}{SST} \qquad SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

* Basic Business Statistic, Berenson et. al.

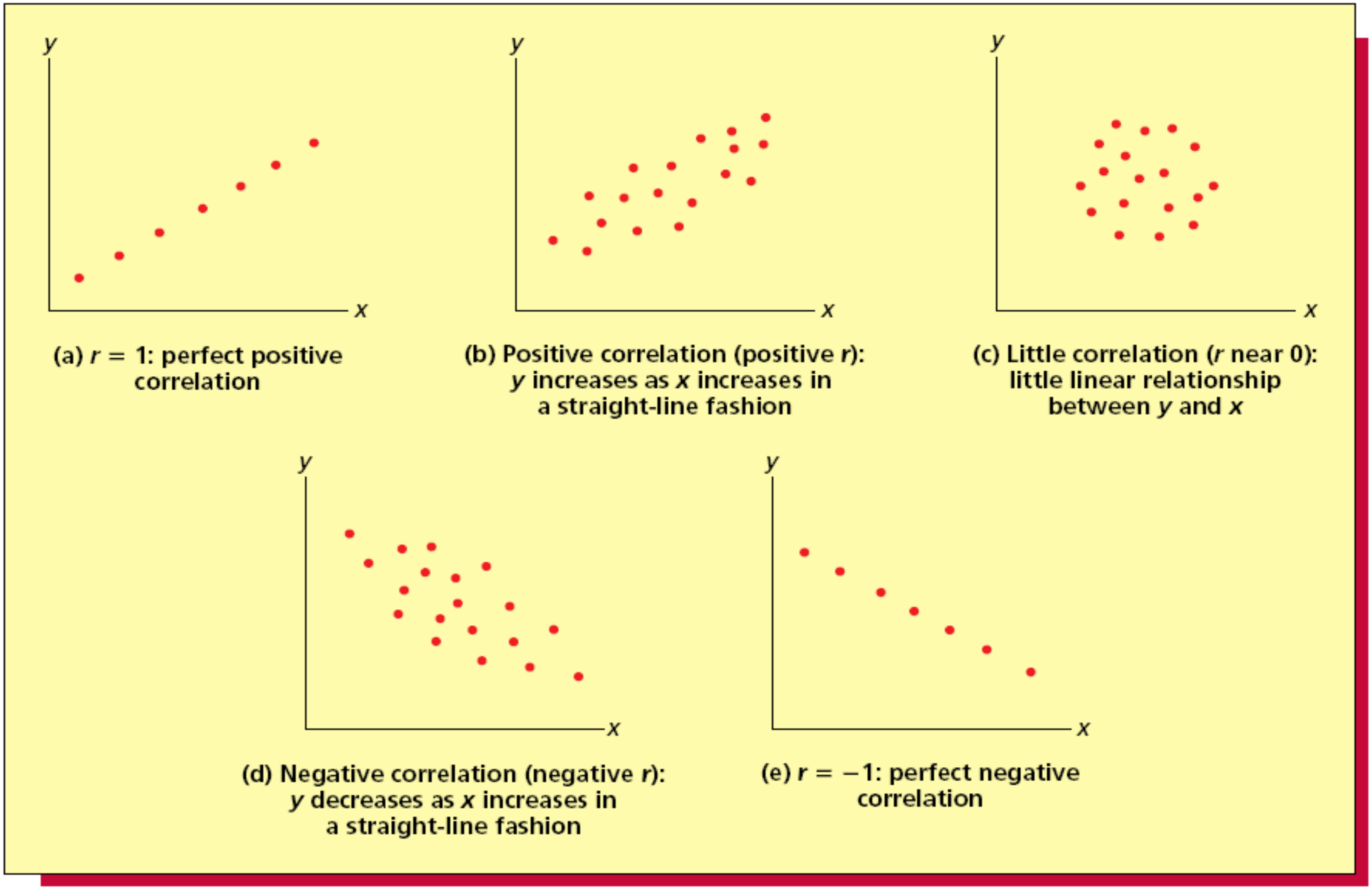
Coefficient of Determination, r^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-square and is denoted as r^2

$$\boxed{r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}}$$

$$\boxed{0 \le r^2 \le 1}$$

Coefficient of Determination, r^2



Regression Coefficients

- The values of the regression parameters β_0, and β_1 are not known. They are estimated from data.

- β_1 indicates the change in the mean response per unit increase in X.

Least Squares Method

- If the scatter plot of our sample data suggests a linear relationship between two variables
  - we can summarize the relationship by drawing a straight line on the plot.

- Least squares method give us the “best” estimated line for our set of sample data.

$$y = \beta_0 + \beta_1 x$$

Regression Line

- We form an estimated regression line based on sample data as

$$\hat{y} = b_0 + b_1 x$$

- The method of least squares chooses the values for b_0, and b_1 to minimize the sum of squared errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y - b_0 - b_1 x\right)^2$$

Least Squares

- Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. Hence errors are squared.

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

- Least Squares minimizes the Sum of the Squared Differences (errors) (SSE)

Regression Line

- We use calculus to derive the minimal values.
- The derived formulas are:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b_1 = r \frac{S_y}{S_x}$$

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

The Least Squares Method

b_0 and b_1 are obtained by finding the values that minimize the sum of the squared differences between the predicted and actual values

$$\boxed{\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2}$$

Project 1

- Part A: Complete the tutorial located at the following URL.
  - https://tutorials.technology/tutorials/19-how-to-do-a-regression-with-sklearn.html
  - You will have to install or have access to Anaconda for this to work.
  - Note, see my email for two key corrections needed.

- Part B: Complete 3.6.2 and 3.6.3 from the tutorial located at the following URL.
  - http://www.scipy-lectures.org/packages/scikit-learn/

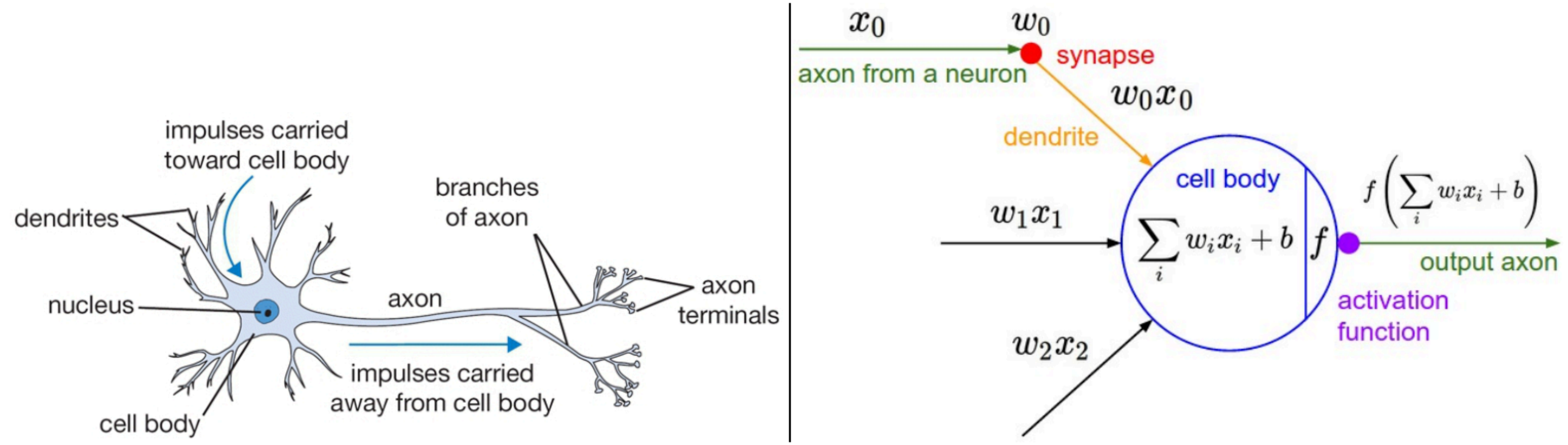
- These two tutorials will introduce you to ML frameworks and will quickly introduce you to key concepts.

The Linear Regression Model

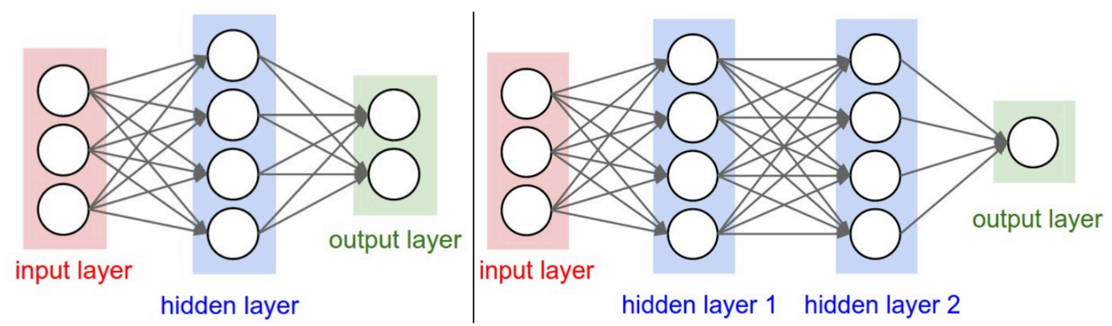
$$\boxed{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon}$$

- Extension to greater than 2 dimensions
- β_0 is the intercept (i.e. the average value for Y if all the X's are zero), β_j is the slope for the jth variable X_j
- β_j is the average increase in Y when X_j is increased by one and **all other X's are held constant.**

Perceptron

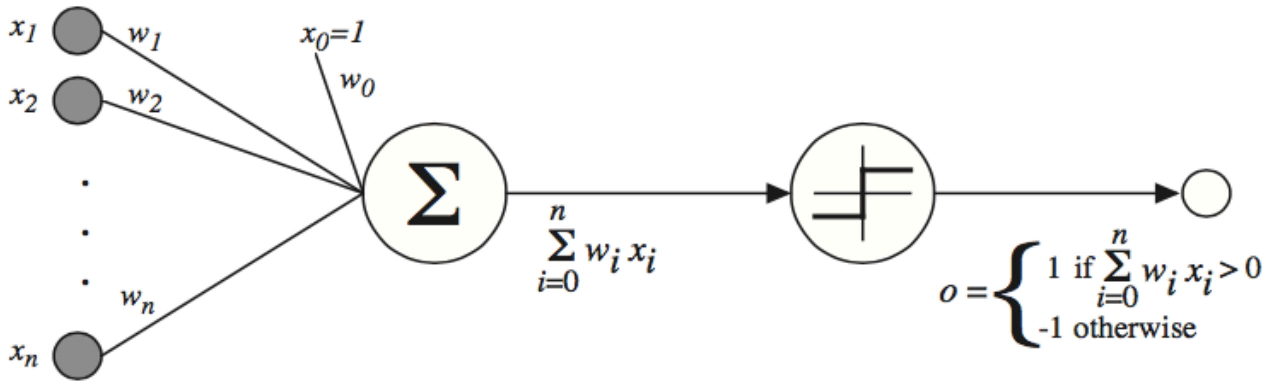


A cartoon drawing of a biological neuron (left) and its mathematical model (right).



Left: A 2-layer Neural Network (one hidden layer of 4 neurons (or units) and one output layer with 2 neurons), and three inputs. **Right:** A 3-layer neural network with three inputs, two hidden layers of 4 neurons each and one output layer. Notice that in both cases there are connections (synapses) between neurons across layers, but not within a layer.

Perceptron



$$o(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Sometimes we'll use simpler vector notation:

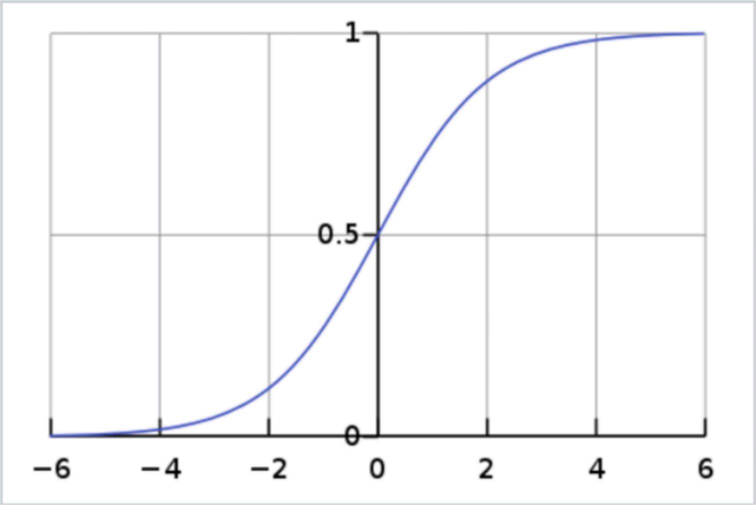
$$o(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{x} > 0 \\ -1 & \text{otherwise.} \end{cases}$$

* T. Mitchell

Sigmoid

- Sigmoid function. Historically famous activation function
- Also known as Inverse logistic
- Works well for usual business related a data
- Works well for a binary or multi class output
- it takes a real-valued number and "squashes" and outputs number between 0 and 1.
- Large negative numbers become 0 and large positive numbers become 1.

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



Rectified Linear Unit ReLU

- Very popular activation function in recent times
- f(x)=max(0,x)
- In other words, the activation is simply thresholder at zero
- Very fast compared to sigmoid and tanH
- Works very well for a certain class of problems
- Doesn't have vanishing gradient problem. It can be used for modelling real values

$f(x)=\max(0,x)$
$f(x)=\log(1+e^x)$