NLP Tutorial: Machine Learning

February 24, 2022

Introduction

Tom Mitchell, 1997

- A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P** if its performance at tasks T, as measured by P, improves with experience E.
- For example Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam mails.
  1. T ?
  2. E ?
  3. P ?
- The sub-field of computer science “gives a computer an ability to learn without being explicitly programmed".

Types of Machine learning Algorithms

- Supervised learning
- Unsupervised learning

Supervised Learning

- Types of learning algorithm in which we have associated truth values for each sample during training.
- Based on the nature of output:
  - **Regression:** A regression problem is when the output variable is a real value, such as predict the age on the basis of the given picture.
  - **Classification:** A classification problem is when the output variable is categorical, such as whether the tumor is malignant or benign.

Linear Regression

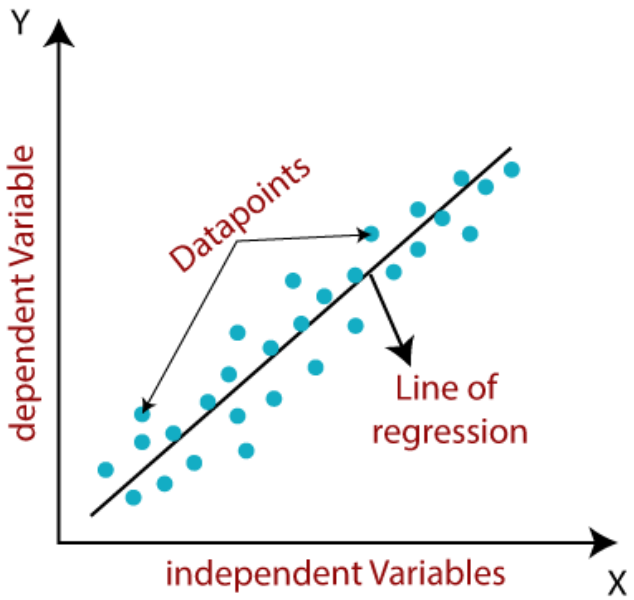
Linear regression is a linear model, that assumes a linear relationship between the input variables (x) and the output variable (y).

	Size in sq. ft. (x)	Price ($) in 1000's (y)
(x^1, y^1)	2104	460
(x^2, y^2)	1416	232
(x^i, y^i)	1534	315
...	...	...
(x^m, y^m)	...	...

Table: Training set for housing problem

- A pair (x^i, y^i) is called a training sample.
- A list of m training samples $\{(x^i, y^i), i = 1, \ldots, m\}$ is called a training set.

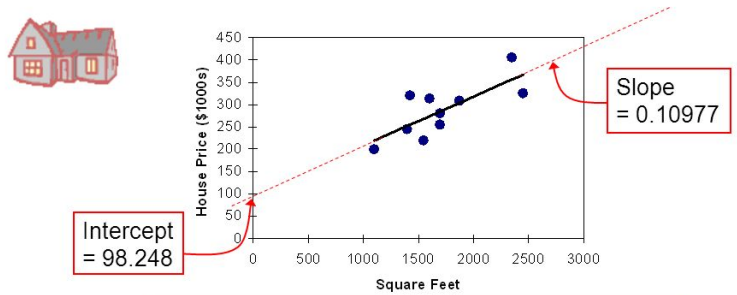
Linear Regression



Simple Linear Regression Example: Graphical Representation

DCOVA

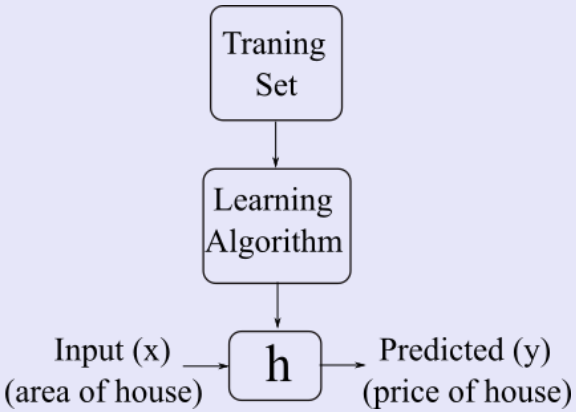
House price model: Scatter Plot and Prediction Line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

https://slideplayer.com/slide/13495114/

Model Representation



- **Hypothesis Function:** $h_\theta(x) = \theta_0 + \theta_1 x$. (where θ's : Parameters)
- How to choose θ_i's?
  - Intuition: Choose θ_0 & θ_i so that $h_\theta(x)$ is close to y for our training sample (x^i, y^i).

Cost Function

- To measure the accuracy of our hypothesis function.
- Also called as Squared error function or Mean squared error.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=0}^{m-1} (h_\theta(x) - y_i)^2 \tag{1}$$

Gradient Decent & Learning Rate

Gradient decent

- To minimize cost function $J(\theta_0, \theta_1)$.
- Gradient descent is an optimization algorithm that enables a model to learn the gradient or direction that the model should take in order to reduce errors.
- repeat until convergence{

  $$\theta_j = \theta_j - \alpha \frac{\delta}{\delta \theta_j} J(\theta_0, \theta_1) \text{ (for j = 0 \& j = 1)}$$

  }
- α is learning rate.
- Both the θ 's are are updated simultaneously.

Logistic Regression

- Logistics Regression is a useful regression method for solving the classification problem.
- It uses the sigmoid function.
- The sigmoid function gives an 'S' shaped curve that can take any real-valued number and map it into a value between 0 and 1.
- Hypothesis function given by:

$$h_\theta(x) = g(z) \tag{2}$$

$$z = \theta^T x \tag{3}$$

$$g(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

Artificial Neural Network (ANN)

- An artificial neural network is a highly interconnected network of large number of processing elements called neuron.
- Neural network derives its origin from human brain:
  - Knowledge is acquired by the network from its environment through a learning process.
  - Inter neuron connection strengths, known as synaptic weights, are used to store the acquired knowledge.
- Non linear classification.

Types of neural network architecture

- Single-Layer Feedforward Networks
- Multi-Layer Feedforward Networks
- Recurrent Neural Networks

Single-Layer Feedforward Networks

- An input layer of source nodes that projects directly onto an output layer of neurons.
- “Single layer" referring to the output layer of computation nodes (neurons).

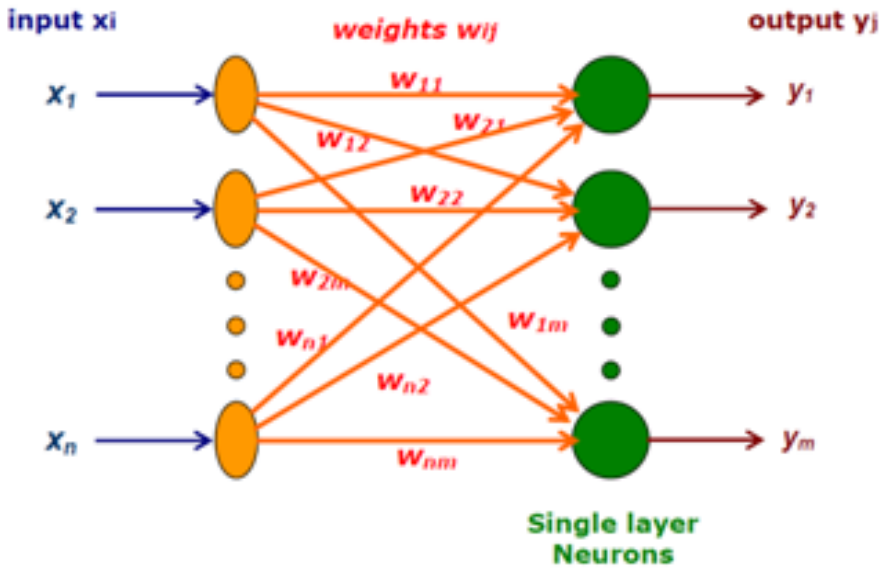


Figure: Single-Layer Feedforward Networks

Multi-Layer Feedforward Networks

- Consists of one or more layer of hidden neurons.
- Hidden layers are responsible for computation.
- More the hidden layer, more the complexity of the network, but efficient output is produced.

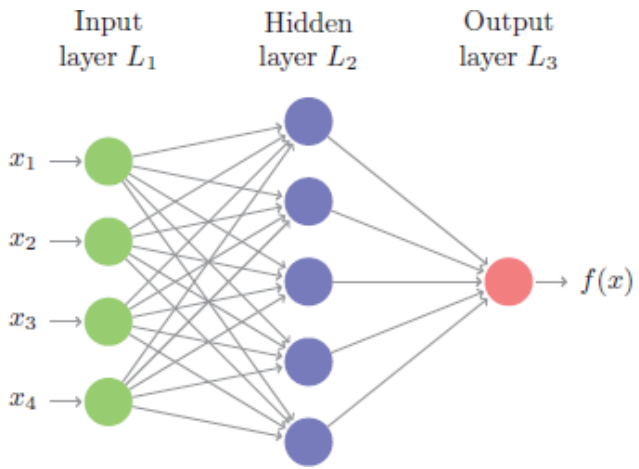


Figure: Multi-Layer Feedforward Networks

Recurrent Neural Networks

- Consist of at least one feedback loop.
- x_t input at time/step t. U, W, V shared across all steps.
- s_t hidden state (memory) at t: $s_t = g(Ux_t + Ws_{t-1})$.
- o_t output at t: $o_t = \text{softmax}(Vs_t)$.

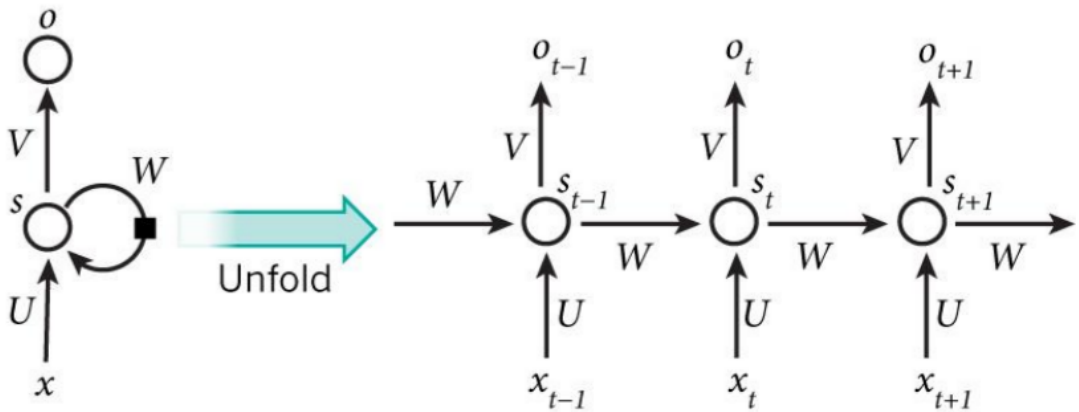


Figure: Recurrent Neural Network.

Nature: tutorial by Danny Britz 2015 @ WILDML

Unsupervised Machine Learning

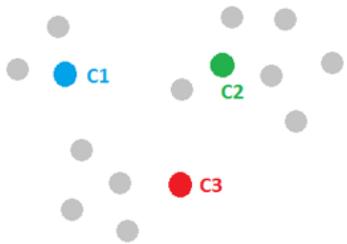
Unsupervised Learning

- In unsupervised learning, we use unlabeled data.
- The goal for unsupervised learning is to model the underlying structure or pattern to learn more about the data.
- Clustering Algorithms are the examples of unsupervised learning.
  - Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data.
  - Document Clustering, Image segmentation, etc.

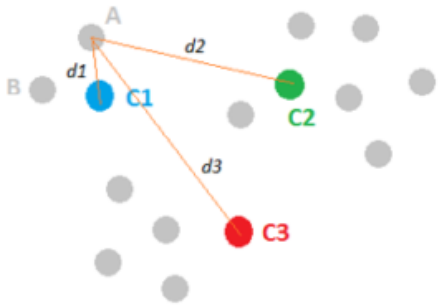
K-Means Clustering Algorithm

- Initialize cluster centers. (randomly pick k points) .
- Assign observations to the closest cluster center.
- Revise cluster centers as mean of assigned observations.
- Repeat step 2 and step 3 until convergence.

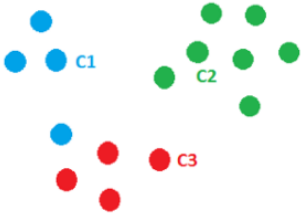
(a)



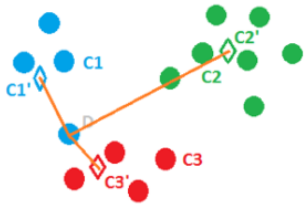
(b)



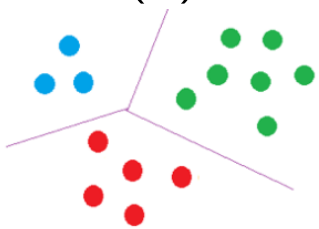
(c)



(d)



(e)



(f)

Underfitting and Overfitting

Overfitting

- The model is performing too well on the training data but the performance drops significantly over the test set.
- This is also known as high variance problem.
- Overfitting occurs when the machine learning model captures the noise from the data.

Underfitting

- The model is performing poorly over the training and the test dataset.
- This is also known as high bias problem.
- Underfitting occurs when a machine learning model cannot capture the underlying trend of the data.

Confusion Matrix

- A confusion matrix is a matrix representation to describe the performance of a classification model.
- The tabular representation between the actual and predicted values as shown below:

		Actual Values	
		Positive	Negative
Predicted Values	Predicted Positive	True Positive (TP)	False Positive (FP)
	Predicted Negative	False Negative (FN)	True Negative (TN)

- **Accuracy:** Accuracy is the percentage of correctly classified cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

- **Precision:** Precision is the percentage of correctly predicted positive observations of the total predicted positive observations.

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{6}$$

- **Recall:** Recall is the percentage of correctly predicted positive observations of all observations in the actual class.

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{7}$$

- **F1-Score:** F1-score is the harmonic mean of precision and recall.

$$F1 - score = \frac{2 \times P \times R}{P + R} \tag{8}$$

Thank you