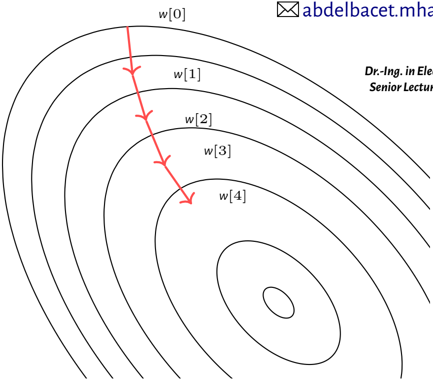
An Introduction To Machine Learning Sorcery

(A DEMYSTIFICATION DRAFT)[1]

Abdelbacet Mhamdi
abdelbacet.mhamdi@bizerte.r-iset.tn

Dr.-Ing. in Electrical Engineering
Senior Lecturer at ISET Bizerte



"Computers are able to see, hear and learn. Welcome to the future."

Dave Waters

"This is nothing. In a few years, that bot will move so fast you'll need a strobe light to see it. Sweet dreams..."

Elon Musk

"Machine intelligence is the last invention that humanity will ever need to make."

Nick Bostrom

[1]Available @ https://github.com/a-mhamdi/isetbz/

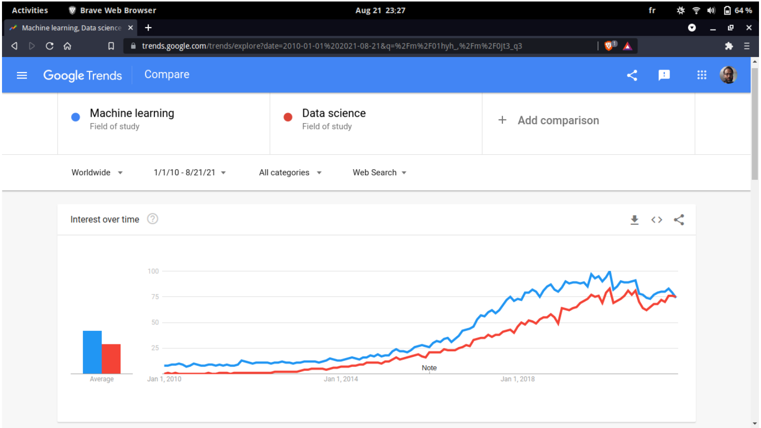
Disclaimer

This document features some material gathered from multiple online sources. Please note no copyright infringement is intended, and I do not own nor claim to own any of the original material. They are used for educational purposes only. I have included links solely as a convenience to the reader. Some links within these slides may lead to other websites, including those operated and maintained by third parties. The presence of such a link does not imply a responsibility for the linked site or an endorsement of the linked site, its operator, or its contents.

Outline

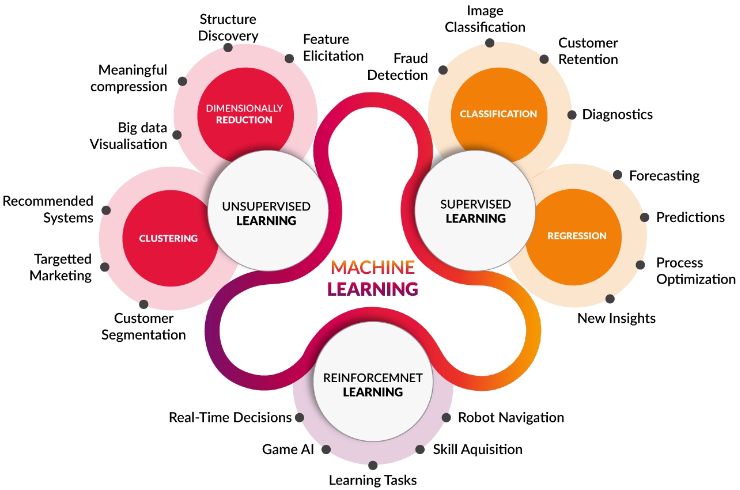
Next...

Top uses

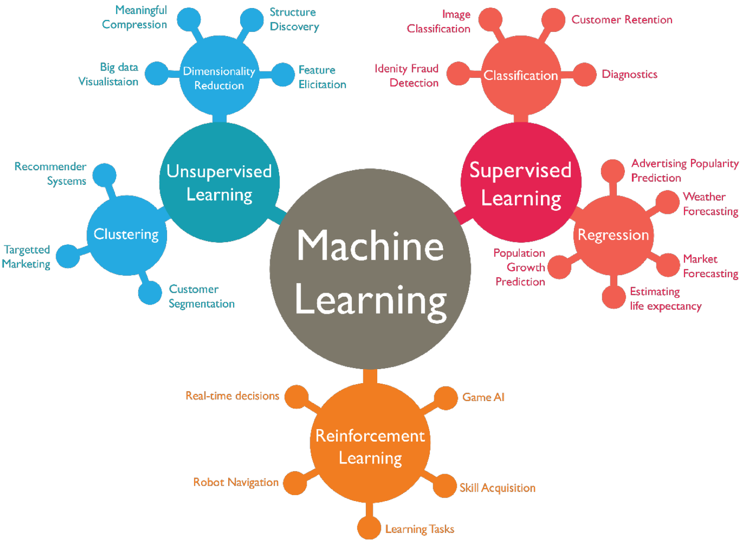


"Numbers represent search interest relative to the highest point on the chart for the given region and time.

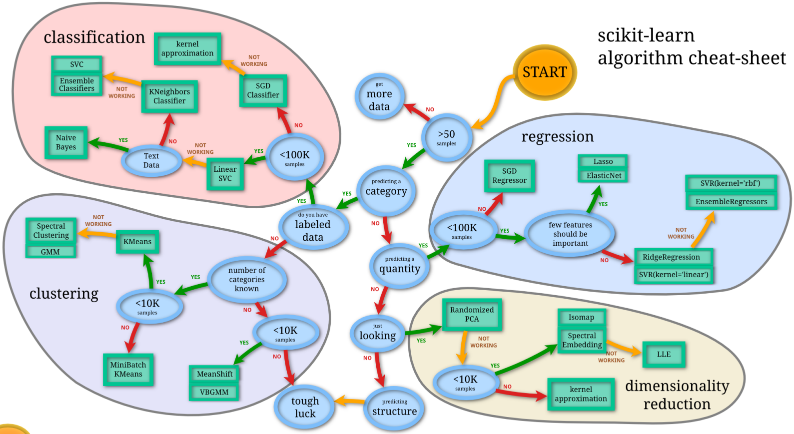
- A value of 100 is the peak popularity for the term;
- A value of 50 means that the term is half as popular;
- A score of 0 means there was not enough data for this term."



https://www.cognub.com/index.php/cognitive-platform/



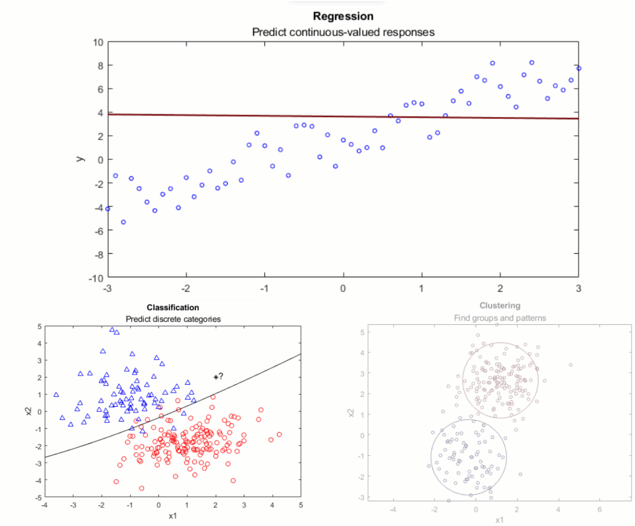
https://vitalflux.com/great-mind-maps-for-learning-machine-learning/



Back

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Regression | Classification | Clustering



https://github.com/MathWorks-Teaching-Resources/Machine-Learning-for-Regression

Machine Learning Definition

Arthur Samuel (1959)

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (1998)

Well-posed Learning Problem: A computer is said to learn from experience $\mathcal{E}$ with respect to some task $\mathcal{T}$ and some performance measure $\mathcal{P}$, if its performance on $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$.

Machine Learning Definition

Arthur Samuel (1959)

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (1998)

Well-posed Learning Problem: A computer is said to learn from experience $\mathcal{E}$ with respect to some task $\mathcal{T}$ and some performance measure $\mathcal{P}$, if its performance on $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$.

Task #1

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. What is the task $\mathcal{T}$ in this setting?

1. Classifying emails as spam or not spam;
2. Watching you label emails as spam or not spam;
3. The number (or fraction) of emails correctly classified as spam/not spam;
4. None of the above-this not a machine learning problem.

Machine Learning Definition

Arthur Samuel (1959)

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (1998)

Well-posed Learning Problem: A computer is said to learn from experience $\mathcal{E}$ with respect to some task $\mathcal{T}$ and some performance measure $\mathcal{P}$, if its performance on $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$.

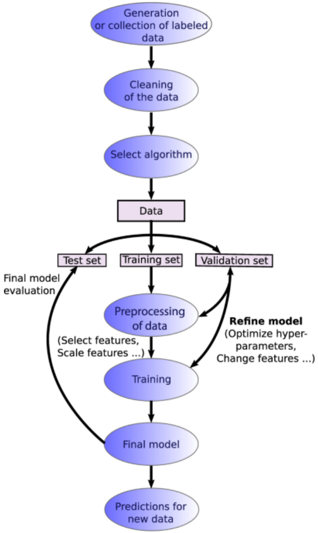
Task #1

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. What is the task $\mathcal{T}$ in this setting?

1. Classifying emails as spam or not spam;
2. Watching you label emails as spam or not spam;
3. The number (or fraction) of emails correctly classified as spam/not spam;
4. None of the above-this not a machine learning problem.

Next...

Overall Methodology [Sch+19]

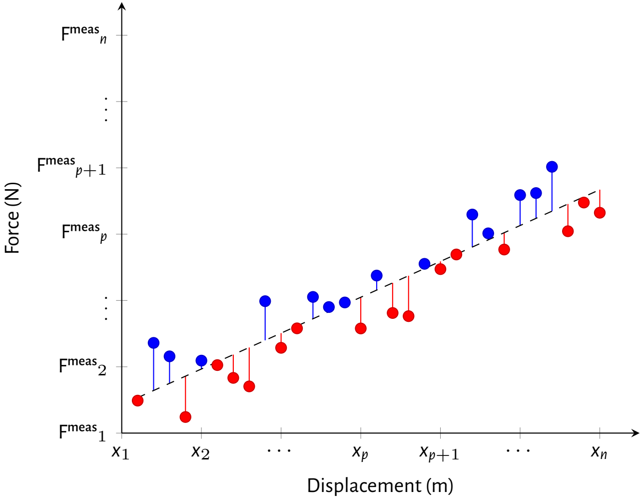


Feature Scaling

Normalisation	Standardisation
$X = \frac{X - \min(X)}{X.\max() - X.\min()} \in [0,\ 1]$	$X = \frac{X - X.\text{mean}()}{X.\text{std}()}$
▲ No assumption on data distribution	▲ More recommended when following normal distribution

[ ]:
```python
from sklearn.preprocessing import StandardScaler
```

[ ]:
```python
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```



Force (N)
Fmeas$_n$
⋮
Fmeas$_{p+1}$
Fmeas$_p$
⋮
Fmeas$_2$
Fmeas$_1$
x_1
x_2
· · ·
x_p
x_{p+1}
· · ·
x_n
Displacement (m)

Consider the example of a spring. Our main goal is to determine the stiffness k of this spring, given some experimental data. The mathematical model *(Hooke's law)*:

$$F = kx \tag{1}$$

Restoring force is proportional to displacement.

Table: Measurements of couple $(x_i,\ \mathrm{F^{meas}}_i)$

x_i	x_1	$\cdots$	x_p	$\cdots$	x_n
$\mathrm{F^{meas}}_i$	$\mathrm{F^{meas}}_1$	$\cdots$	$\mathrm{F^{meas}}_p$	$\cdots$	$\mathrm{F^{meas}}_n$

$$\begin{aligned} \mathrm{F^{meas}}_i &= \mathrm{F}_i + \varepsilon_i \\ &= kx_i + \varepsilon_i, \end{aligned} \tag{2}$$

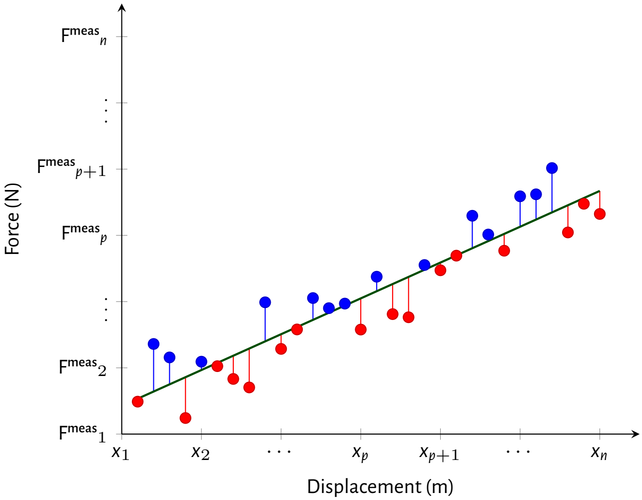
where F_i denotes the unknown real value of the force applied to the spring. In order to estimate the stiffness value k, we can consider the quadratic criterion:

$$\begin{aligned} \mathcal{J} &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} (\mathrm{F^{meas}}_i - kx_i)^2 \end{aligned}$$

$$\frac{\partial \mathcal{J}}{\partial k} = 0 \tag{3}$$

$$\begin{aligned} 2\sum_{i=1}^{n}\left(\mathsf{F^{meas}}_i - kx_i\right)\sum_{i=1}^{n}\frac{\partial\left(\mathsf{F^{meas}}_i - kx_i\right)}{\partial k} &= 0 \\ \sum_{i=1}^{n}\left(\mathsf{F^{meas}}_i - kx_i\right)\sum_{i=1}^{n}x_i &= 0 \end{aligned}$$

$$\sum_{i=1}^{n}\mathsf{F^{meas}}_i\, x_i = k\sum_{i=1}^{n}x_i^2 \iff \boxed{\hat{k} = \frac{\sum_{i=1}^{n}\mathsf{F^{meas}}_i\, x_i}{\sum_{i=1}^{n}x_i^2}}$$


$F^{meas}{}_n$
$\vdots$
$F^{meas}{}_{p+1}$
$F^{meas}{}_p$
$\vdots$
$F^{meas}{}_2$
$F^{meas}{}_1$
Force (N)
x_1
x_2
$\cdots$
x_p
x_{p+1}
$\cdots$
x_n
Displacement (m)

This example consists on determining the unknown couple $(y_0,\ v_0)$ of a mobile solid. We assume that the trajectory is linear. The mathematical model that relates the position y to time t is given by this equation:

$$y = y_0 + v_0 t \tag{4}$$

Table: Measurements of position y

t_i	t_1	$\cdots$	t_p	$\cdots$	t_n
$y^{meas}{}_i$	$y^{meas}{}_1$	$\cdots$	$y^{meas}{}_p$	$\cdots$	$y^{meas}{}_n$

$$\begin{aligned} y^{meas}{}_i &= y_i + \varepsilon_i \\ &= y_0 + v_0 t_i + \varepsilon_i, \end{aligned} \tag{5}$$

where y_i denotes the unknown real value of the position y at time point t_i.
In order to estimate the values taken by the couple $\begin{bmatrix} y_0, & v_0 \end{bmatrix}^T$, we consider the quadratic criterion again, as follows:

$$\begin{aligned} \mathcal{J} &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \varepsilon^T \times \varepsilon \end{aligned}$$

The vector ε is set by $\varepsilon_i, \ \forall i \geq 1$:

$$\varepsilon = \begin{bmatrix} \varepsilon_1 & \cdots & \varepsilon_n \end{bmatrix}^T$$

$$\frac{\partial \mathcal{J}}{\partial \begin{bmatrix} y_0 \\ \\ v_0 \end{bmatrix}} = 0 \tag{6}$$

Consider the following multivariate equation:

$$y = \theta_1 x_{(1)} + \theta_2 x_{(2)} + \cdots + \theta_m x_{(m)} \tag{7}$$

For a particular single measurement, eq. (7) can be updated as

$$y_k = \theta_1 x_{(1,\,k)} + \theta_2 x_{(2,\,k)} + \cdots + \theta_m x_{(m,\,k)} + \varepsilon_k \tag{8}$$

We denote hereafter by θ the vector $\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$. The function y_k becomes:

$$y_k = \underbrace{\left[x_{(1,\,k)},\, x_{(2,\,k)},\, \cdots,\, x_{(m,\,k)}\right]}_{x_k^T} \theta + \varepsilon_k$$

We assume that we have n measurements for y. Then we can transform the previous equation into

$$y = H\theta + \varepsilon,$$

where $y^T = [y_1,\, y_2,\, \cdots,\, y_n]$, $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$, and $\varepsilon^T = [\varepsilon_1,\, \varepsilon_2,\, \cdots,\, \varepsilon_n]$.

We can consider the mean squared error or quadratic criterion in order to compute the approximated value of θ:

$$\begin{aligned}\mathcal{J} &= \sum_{k=1}^{n} \varepsilon_k^2 \\ &= \varepsilon^T \varepsilon\end{aligned}$$

The best well estimated value of $\hat{\theta}$ corresponds to the absolute minimum of $\mathcal{J}$. This leads to calculate the gradient of $\mathcal{J}$ with respect to θ:

$$\frac{\partial \mathcal{J}}{\partial \theta} = \frac{\partial \left(\varepsilon^T \varepsilon\right)}{\partial \theta} \tag{9}$$

$$\frac{\partial \left(\varepsilon^T \varepsilon\right)}{\partial \theta} = 2 \left(\frac{\partial \varepsilon}{\partial \theta}\right)^T \varepsilon \tag{10}$$

Recall that $\varepsilon = y - X\theta$, the term $\frac{\partial \varepsilon}{\partial \theta}$ hence becomes:

$$\frac{\partial \varepsilon}{\partial \theta} = -X \tag{11}$$

$$\begin{aligned}\frac{\partial J}{\partial \theta} &= 2(-X)^T(y - X\theta)\\ &= 0\end{aligned}$$

The regressor is given by

$$\boxed{\hat{\theta} = (X^T X)^{-1} X^T y}$$

$X^T X$ **is not invertible (singular/degenerate)**

▼ Redundant Features

Some features are linearly dependant, i.e, $\exists$ some $x_p \propto$ some x_l for instance x_p in feet and x_l in m.

▼ Too many features

Fewer observations compared to the number of features, i.e, $m \geq n$.

▲ Delete some features
▲ Add extra observations
▲ Use regularization

Gradient Descent

$$\boxed{\theta_i = \theta_i - \underbrace{\alpha}_{\text{Learning Rate}} \frac{\partial \mathcal{J}}{\partial \theta_i}}$$

Recall that $\mathcal{J} = {}^1/_{2n} \sum_{k=1}^{n} (y_k - h_\theta(x_k))^2 \implies \frac{\partial \mathcal{J}}{\partial \theta_i} = -{}^1/_n \sum_{k=1}^{n} (y_k - h_\theta(x_k))\, x_{(i,\,k)}$

$$\boxed{\theta_i \leftarrow \theta_i + \alpha \frac{1}{n} \sum_{k=1}^{n} (y_k - h_\theta\,(x_k))\, x_{(i,\,k)}}$$

$$\boxed{\theta_0 \leftarrow \theta_0 + \alpha \frac{1}{n} \sum_{k=1}^{n} (y_k - h_\theta\,(x_k))\, x_{(0,\,k)}} \qquad \boxed{\theta_1 \leftarrow \theta_1 + \alpha \frac{1}{n} \sum_{k=1}^{n} (y_k - h_\theta\,(x_k))\, x_{(1,\,k)}}$$

$$\vdots$$

$$\boxed{\theta_m \leftarrow \theta_m + \alpha \frac{1}{n} \sum_{k=1}^{n} (y_k - h_\theta\,(x_k))\, x_{(m,\,k)}}$$

Task #2

The yield y of a chemical process is a random variable whose value is considered to be a linear function of the temperature x. The following data of corresponding values of x and y is found:

Temperature in °C (x)	0	25	50	75	100
Yield in grams (y)	14	38	54	76	95

The linear regression model $y = \theta_0 + \theta_1 x$ is used. Determine the values of $\theta_0,\ \theta_1$.

1. Using normal equation,
2. Using gradient descent for 5 iterations.

$$y = \begin{bmatrix} 14 \\ 38 \\ 54 \\ 76 \\ 95 \end{bmatrix} \quad \text{et} \quad X = \begin{bmatrix} 1 & 0 \\ 1 & 25 \\ 1 & 50 \\ 1 & 75 \\ 1 & 100 \end{bmatrix} \Longrightarrow X^T X = \begin{bmatrix} 5 & 250 \\ 250 & 18750 \end{bmatrix}$$

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} = \begin{bmatrix} 15.4 \\ 0.8 \end{bmatrix}$$

?? PythonTeX ??

F1-Score, **Accuracy**, **Recall** and **Precision** are calculated as follow:

$$f1 - \text{score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

$f1$ − score denotes the *Harmonic Mean of Recall & Precision*

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

It denotes the ratio of how much we got right over all cases. Recall, on the other hand, designates the ratio of how much positives do we got right over all actual positive cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

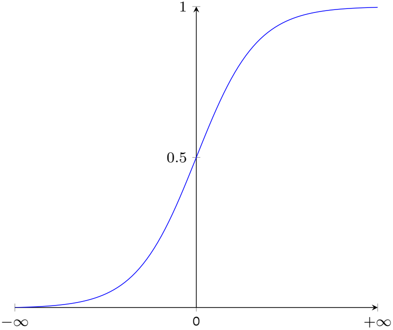
Precision, at last, is how much positives we got right over all positive predictions. It is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Method	Pros		Cons	
Logistic Regression	▲	Probabilistic	▼	almost linearly separable data
K-NN	▲ ▲ ▲	Simple Fast Efficient	▼	Number of neighbors k
SVM	▲ ▲	*** ***	▼ ▼	*** ***
Kernel SVM	▲ ▲	*** ***	▼ ▼	*** ***
Naive Bayes	▲ ▲	*** ***	▼ ▼	*** ***
Decision Tree Classification	▲ ▲	*** ***	▼ ▼	*** ***
Random Forest Classification	▲ ▲	*** ***	▼ ▼	*** ***

Logistic or S-shaped function σ

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



- σ squashes range of distance from $]-\infty, +\infty[$ to $[0, 1]$
- σ is differentiable and easy to compute:
  $\boxed{\dot{\sigma} = \sigma \times (1 - \sigma)}$

Decision boundary

$$y = \sigma\left(\theta_1 x_{(1)} + \theta_2 x_{(2)} + \cdots + \theta_m x_{(m)}\right)$$

$$y = \frac{1}{1 + \mathrm{e}^{-\theta^T x}}$$

Hypothesis:

$$h_\theta\left(x\right) = \frac{1}{1 + \mathrm{e}^{-\theta^T x}} \qquad h_\theta\left(x_k\right) = \frac{1}{1 + \mathrm{e}^{-\theta^T x_k}}$$

Cost function:

$$\mathcal{J} = \begin{cases} -\ln\left(h_\theta(x)\right) & \text{if} \quad y = 1 \\ -\ln\left(1 - h_\theta(x)\right) & \text{if} \quad y = 0 \end{cases}$$

$$\boxed{\mathcal{J} = -y \ln\left(h_\theta(x)\right) - (1 - y)\ln\left(1 - h_\theta(x)\right)}$$

Gradient Descent

$$\boxed{\theta_i = \theta_i - \underbrace{\alpha}_{\text{LEARNING RATE}} \frac{\partial \mathcal{J}}{\partial \theta_i}}$$

Generalizing $\mathcal{J}$ yields: $\mathcal{J} = -\frac{1}{n}\sum_{k=1}^{n}\left(y_k \ln\left(h_\theta(x_k)\right) + (1-y_k)\ln\left(1-h_\theta(x_k)\right)\right)$

$$\implies \quad \frac{\partial \mathcal{J}}{\partial \theta_i} = -\frac{1}{n}\sum_{k=1}^{n}\left(y_k - h_\theta(x_k)\right) x_{(i,\,k)}$$

$$\boxed{\theta_i \leftarrow \theta_i + \alpha\frac{1}{n}\sum_{k=1}^{n}\left(y_k - h_\theta\left(x_k\right)\right) x_{(i,\,k)}}$$

$$\boxed{\theta_0 \leftarrow \theta_0 + \alpha\frac{1}{n}\sum_{k=1}^{n}\left(y_k - h_\theta\left(x_k\right)\right) x_{(0,\,k)}} \qquad \boxed{\theta_1 \leftarrow \theta_1 + \alpha\frac{1}{n}\sum_{k=1}^{n}\left(y_k - h_\theta\left(x_k\right)\right) x_{(1,\,k)}}$$

$$\vdots$$

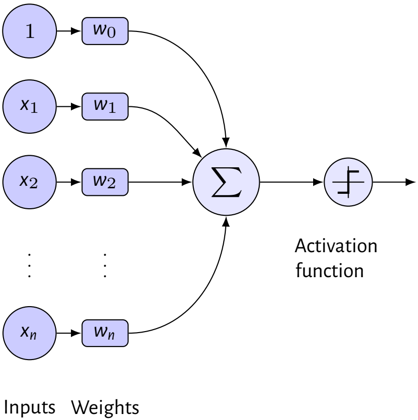
$$\boxed{\theta_m \leftarrow \theta_m + \alpha\frac{1}{n}\sum_{k=1}^{n}\left(y_k - h_\theta\left(x_k\right)\right) x_{(m,\,k)}}$$

Next...

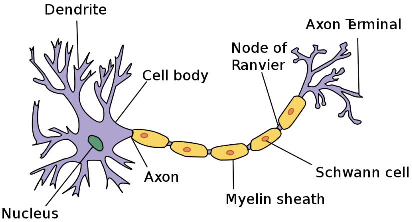
Next...

Fundamental unit of a neural network (1/2)

Artificial neuron

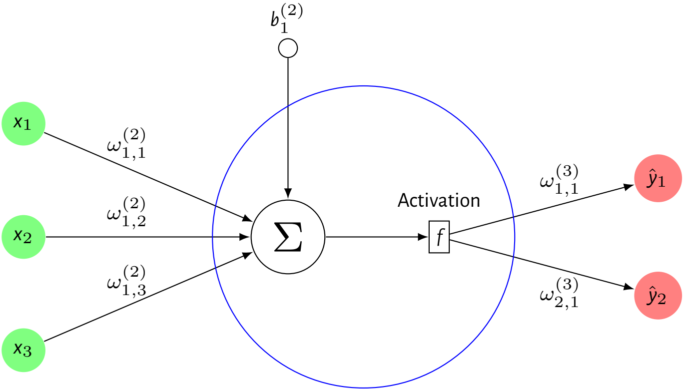


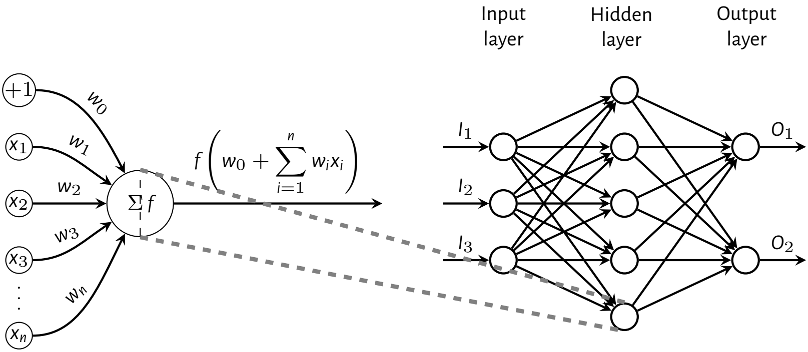
Biological neuron



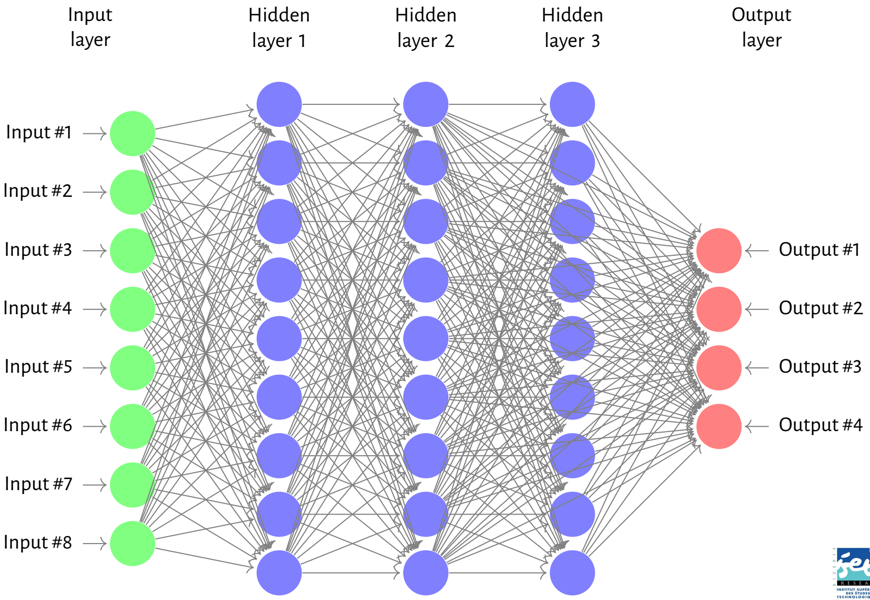
https://id.wikipedia.org/wiki/Sel_saraf

Fundamental unit of a neural network (2/2)




+1
x_1
x_2
x_3
x_n
w_0
w_1
w_2
w_3
w_n
$\Sigma\ f$
$f\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$
Input layer
Hidden layer
Output layer
I_1
I_2
I_3
O_1
O_2

Multilayer Perceptron (MLP)



Next...

MCQ (1/14)

Features of Machine Learning are ...

1. Automation
2. Improved customer experience
3. Business intelligence
4. All of the above

The term machine learning was coined by ...

1. James Gosling
2. Arthur Samuel
3. Guido van Rossum
4. None of the above

Which among the following algorithms are used in Machine learning?

1. Naive Bayes
2. Support Vector Machines
3. K-Nearest Neighbors
4. All of the above

MCQ (2/14)

... is the machine learning algorithm that can be used with labeled data.

1. Regression algorithm
2. Clustering algorithm
3. Association algorithm
4. All of the above

Replace missing values with mean/median/mode helps to handle missing or corrupted data in a dataset. True/False?

1. True
2. False

The Real-world machine learning use cases are

1. Digital assistants
2. Chatbots
3. Fraud detection
4. All of the above

MCQ (3/14)

... is a part of machine learning that works with neural networks.

1. Deep learning
2. Artificial intelligence
3. All of the above
4. None of the above

The supervised learning problems can be grouped as ...

1. Regression problems
2. Classification problems
3. All of the above
4. None of the above

The unsupervised learning problems can be grouped as ...

1. Clustering
2. Association
3. All of the above
4. None of the above

MCQ (4/14)

Overfitting is a type of modelling error which results in the failure to predict future observations effectively or fit additional data in the existing model. Yes/No?

1. Yes
2. No
3. Can not say
4. Probably

... is the scenario when the model fails to decipher the underlying trend in the input data.

1. Underfitting
2. Overfitting
3. All of the above
4. None of the above

MCQ (5/14)

Machine learning approaches can be traditionally categorized into ... categories.

1. 3
2. 4
3. 7
4. 9

The categories in which Machine learning approaches can be traditionally categorized are ...

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning
4. All of the above

MCQ (6/14)

In general, to have a well-defined learning problem, we must identity which of the following

1. The class of tasks
2. The measure of performance to be improved
3. The source of experience
4. All of the above

The average positive difference between computed and desired outcome values

1. Root Mean Squared Error
2. Mean Squared Error
3. Mean Absolute Error
4. Mean Positive Error

... is used as an input to the machine learning model for training and prediction purposes.

1. Target variable
2. Feature vector
3. All of the above
4. None of the above

MCQ (7/14)

Simple regression assumes a ... relationship between the input attribute and output attribute.

1. linear
2. quadratic
3. reciprocal
4. inverse

What is Machine Learning (ML)?

1. The selective acquisition of knowledge through the use of computer programs
2. The selective acquisition of knowledge through the use of manual programs
3. The autonomous acquisition of knowledge through the use of computer programs
4. The autonomous acquisition of knowledge through the use of manual programs

MCQ (8/14)

The correlation between the number of years an emplyee has worked for a company and the salary of the employee is 0.75. What can be said about emplyee salary and years worked?

1. There is no relationship between salary and years worked.
2. Individuals that have worked for the company the longest have higher salaries.
3. Individuals that have worked for the company the longest have lower salaries.
4. The majority of employees have been with the company a long time.

Successful applications of ML

1. Learning to recognize spoken words
2. Learning to drive an autonomous vehicle
3. Learning to classify new astronomical structures
4. Learning to play world-class backgammon
5. All of the above

MCQ (9/14)

Which machine learning models are trained to make a series of decisions based on the rewards and feedback they receive for their actions?

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning
4. All of the above

Which of the following is not a type of supervised learning?

1. Classification
2. Regression
3. Clustering
4. None of the above

MCQ (10/14)

As the amount of training data increases

1. Training error usually increases and generalization error usually increases
2. Training error usually increases and generalization error usually decreases
3. Training error usually decreases and generalization error usually decreases
4. Training error usually decreases and generalization error usually increases

Which of the following are not classification tasks?

1. Find the gender of a person by analyzing his writing style
2. Detect Pneumonia from Chest X-ray images
3. Predict the price of a house based on floor area, number of rooms, etc.
4. Predict whether there will be abnormally heavy rainfall next year

Which of the following is a categorical feature?

1. Height of a person
2. Price of petroleum
3. Amount of rainfall in a day
4. Mother tongue of a person

MCQ (11/14)

What is the use of validation dataset in Machine Learning?

1. To train the machine learning model.
2. To tune the hyperparameters of the machine learning model
3. To evaluate the performance of the machine learning model
4. None of the above

Which of the following criteria is typically used for optimizing in linear regression.

1. Maximize the number of points it touches.
2. Minimize the number of points it touches.
3. Minimize the squared distance from the points.
4. Minimize the maximum distance of a point from a line.

For two runs of K-Mean clustering, is it expected to get same clustering results?

1. Yes
2. No

MCQ (12/14)

Which of the following can act as possible termination conditions in K-Means?

a) For a fixed number of iterations

b) Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.

c) Centroids do not change between successive iterations

d) Terminate when RSS falls below a threshold

1. a, c & d
2. a, b & c
3. a, b & d
4. All of the above

In training a neural network, we notice that the loss does not increase in the first few starting epochs: What is the reason for this?

1. The learning rate is low
2. Regularization parameter is high
3. Stuck at the local minima
4. All of the above

MCQ (13/14)

Which of the following is true about model capacity (*where model capacity means the ability of neural network to approximate complex functions*)?

1. As number of hidden layers increases, model capacity increases
2. As dropout ratio increases, model capacity increases
3. As learning rate increases, model capacity increases
4. None of these

When there is noise in data, which of the following options would improve the performance of the KNN algorithm?

1. Increase the value of k
2. Decrease the value of k
3. Changing value of k will not change the effect of the noise
4. None of these

MCQ (14/14)

Logistic Regression is used for ...

1. regression purposes
2. classification purposes
3. all of the above
4. none of the above

Which of the following methods do we use to best fit the data in Logistic Regression?

1. Least Squared Error
2. Maximum Likelihood
3. Jaccard distance

Further Reading

[ENM15] I. El Naqa and M. J. Murphy. "What Is Machine Learning?" In: *Machine Learning in Radiation Oncology: Theory and Applications*. Ed. by I. El Naqa, R. Li, and M. J. Murphy. Cham: Springer International Publishing, 2015, pp. 3–11. DOI: 10.1007/978-3-319-18305-3_1.

[Sch+19] J. Schmidt et al. "Recent advances and applications of machine learning in solid-state materials science". In: *npj Computational Materials* 5.1 (Aug. 2019). DOI: 10.1038/s41524-019-0221-0 (cit. on p. 15).