TRAIN, FIT

Given a statistical model, fit a model means to find the best parameters (trained by the model on its own) that provide a prediction values closer to the real values. Below, an example by house prices data set:

ROW, OBSERVATION, SAMPLE, EXAMPLE, INSTANCE, RECORD

An instance is a row of a data set. In this record you can find values of one or more features and possibly the value of a label. Below, an example by house prices data set:

RIDGE REGULARIZATION, L2 REGULARIZATION, RIDGE REGRESSION

Ridge regression is a form of regression, that constrains/regularizes or shrinks the coefficient estimates towards zero. It’s a technique used to avoid overfitting and it afford a better generalization of the model. Ridge regression is similar to the least squares, except that parameters are estimated by minimizing a different loss function adding a penalty quantity:

The residual sum of squares is modified adding a shrinkage penalty making the residual sum of squares small and , that is the tuning parameter, decides how much to penalize the flexibility of the model.

With , the penalty term has no eﬀect and the estimates produced by ridge regression will be equal to the least squares. With the impact of the shrinkage penalty grows and the ridge regression coeﬃcient estimates will approach towards zero. The penalty term will shrink all the coefficients towards zero, but not set any of them exactly to zero (unless ).

PARTICIPATION BIAS, NON-RESPONSE BIAS

Non-response bias is an event in which results of surveys become non-represantative. Non response happens when there is a significant difference between those who responded at the survey and those who didn’t.

It’s introduced when respondents differ from non-respondents, so the sample is systematically different from the target population, potentially resulting in biased estimates.

Examples of causes: poorly constructed surveys, surveys didn’t reach all members in the reference sample, certain groups were more inclined to answer, some people refused to participate…

MULTINOMIAL CLASSIFICATION, MULTI-CLASS CLASSIFICATION

Multi-class classification is a classification task with more than two classes in the response variable. For example, the response variable to predict is the quality of a product with the follow 4 classes: excellent, good, fair, bad. Instead, an example of binary classification task is to evaluates email messages and the classification either “spam” or “not spam”.

FEATURE EXTRACTION, FEATURE ENGINEERING

<https://stackoverflow.com/questions/39130600/what-is-the-difference-between-feature-engineering-and-feature-extraction>

Feature engineering transform raw data into features/attributes that better represent the underlying structure of the data, they are encoded in a manner that makes it as easy as possible for the model to achieve good performance. Sometimes is called feature extraction.

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. Example: PCA, ICA, t-SNE…

DENSE LAYER, FULLY CONNECTED LAYER

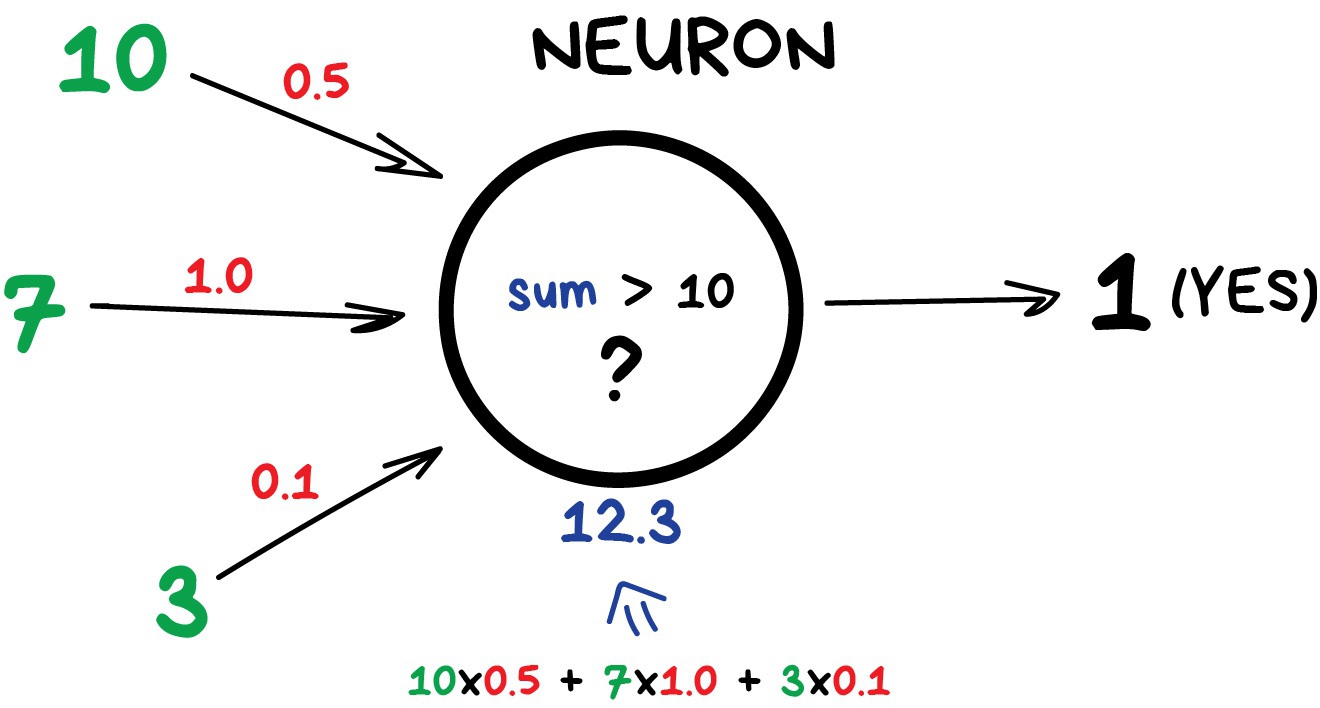
The deep learning architecture is based on many layers and the term "fully connected layer" means that in every hidden layer each node is connected to every node of the next hidden layer.

**DEEP MODEL DEEP NEURAL NETWORK**

Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

There are two types of neural networks: shallow neural networks which have one hidden layer, and deep neural networks which have more than one hidden layer.

Artificial Neural Networks are inspired by neuroscience, they are formed by neurons connected in various ways. Neuron is the basic element block of the artificial neural network. The goal is to create a machine able to learn, able to mimic the human brain, why? Because brain is really powerful tool to learn. Neuron alone is useless, but with others is powerful.

[](https://i.vas3k.ru/7wf.jpg)

Artificial Neural Networks have an input layer with as many neurons as the number of explanatory variables. The input layer is followed by one or several hidden layers with an optional number of neurons. The number of neurons in the output layer equals the number of dependent variables.

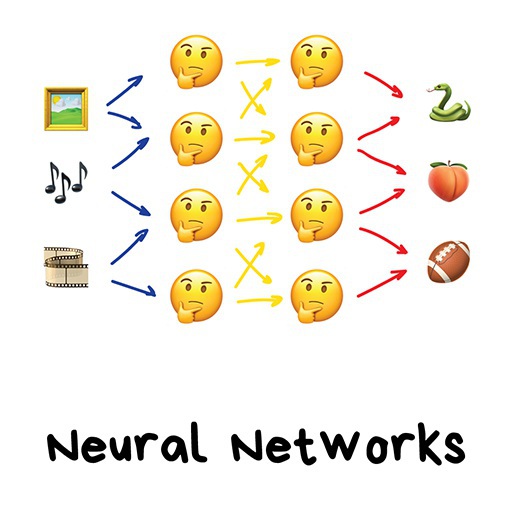
Each neuron or node has input signals and one output signal. Input signals can be features or can be represented by others neurons. Links between input features and neuron are called sinapses and for all these linkes are given weights, they are crucial because weights are how neural networks learn, with weights network decide which signal is relevant or not, which one is transmitted and which one not. Each neuron receives input from each neuron in the preceding layer and feeds its output to every neuron in the next layer.

They are called networks because they are a combination of different functions, so they extend linear models applying a non linear transformation, called activaction function, to the linear combination of the inputs coming from preceding layers.

So in every layer there is a linear combination of weights and input features or latent features coming from preceding layers and then is applied an activaction function.

The neural networks are called feed-forward because the signals propagate from one layer to the next.

The weights in a neural networks are learned using backpropagation, which is an iterative 2-stage supervised learning algorithm. The weights are usually initialized by randomization. In the forward propagation stage, the training data is inputed to the network. The output from the model is compared to the corresponding observed value of the dependent variable. In the backpropagation stage, the weights are adjusted so that the error on the training set is lowered.



DECISION THRESHOLD, CLASSIFICATION THRESHOLD

It’s a numerical value threshold applied to the score predicted by a model to split values belonging to a positive class from values belonging to the negative class in a binary classification.

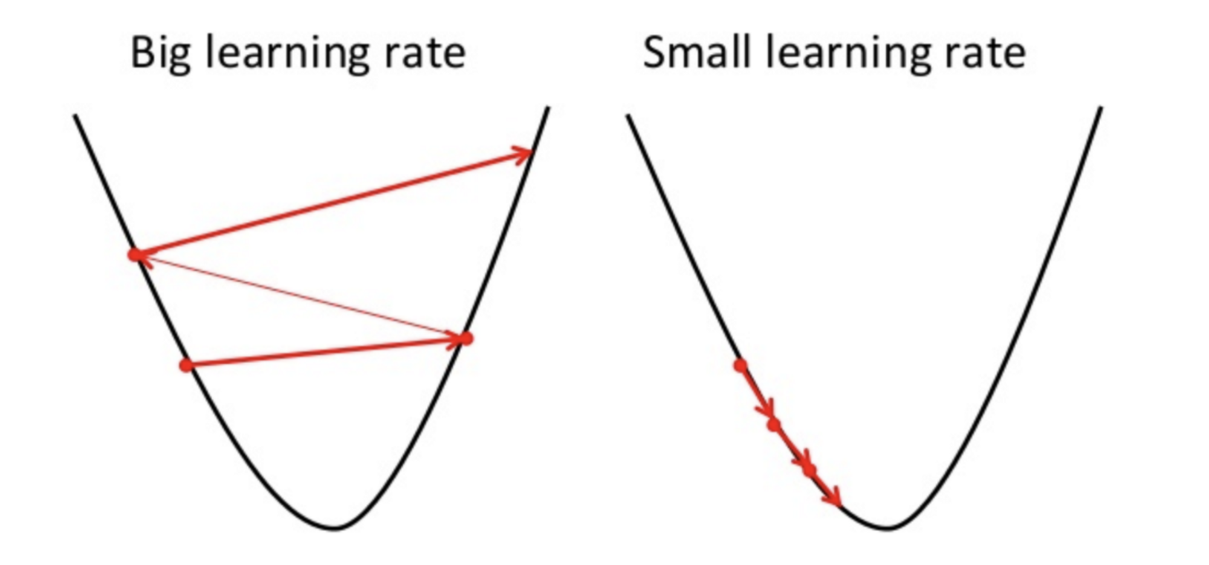
For example, in a email spam classification task if the threshold is 0.8, the values predicted greater than 0.8 will be classified spam, the others will be classified as not spam.

STEP SIZE, LEARNING RATE

Gradient Descent is used while training a machine learning model. It is an optimization algorithm, based on a convex function, that tweaks it’s parameters iteratively to minimize a given function to its local minimum.

Gradient Descent can be thought of climbing down to the bottom of a valley, instead of climbing up a hill.

It starts with initial small random values of the parameters, calculates the partial derivative of the cost function respect the parameters and updates them with small steps (learning rate ) in the opposite direction (sign) of the calculated gradient. The process is repeated until is reached a minimum of the loss function. Learning rate defines how much parameters should change in each iteration. In other words it controls how fast or slow we should converge to the minimum.

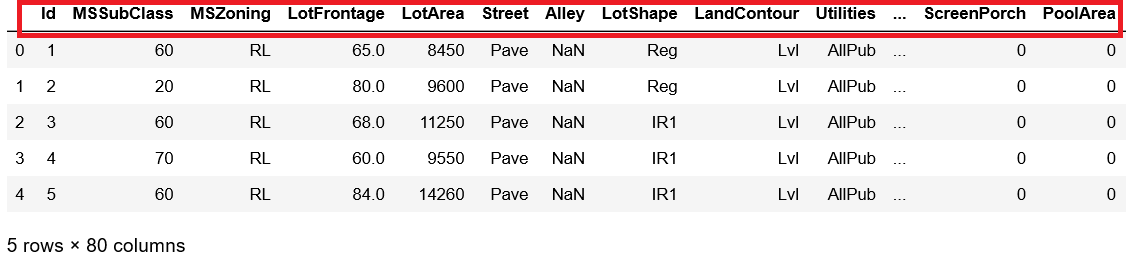


COST, LOSS

It’s a function that measures the performance of a Machine Learning model for given data. Cost Function quantifies the error between predicted values and actual values. It figure out how to fit the best hypothesis function to the observed data. It’s used in the parameter estimation to minimize the distance between the predicted values and the observed values. Examples are MSE, RMSE….

ATTRIBUTE, FEATURE, INDEPENDENT VARIABLE, INPUTS, COLUMN, PREDICTOR, REGRESSOR

A feature is a measurable property or characteristic of an object in a data set that is being used. Features appear as columns and they are the inputs data used in the machine learning model for the prediction. Below, an example by house prices data set:



LABEL, TARGET, DEPENDENT VARIABLE, RESPONSE, OUTPUT, OUTCOME

Label is the output, is the result coming from the prediction of the features. In a Supervised Learning task the goal to achieve with a Machine Learning model is to find the true relationship between the inputs and the outcome. Below, an example by house prices data set:

