SUBJECT CODE : 410251

As per Revised Syllabus of

SAVITRIBAI PHULE PUNE UNIVERSITY

Choice Based Credit System (CBCS)

B.E. (Computer) Semester : VIII

DEEP LEARNING

Iresh A. Dhotre

M.E. (Information Technology)

Ex-Faculty, Sinhgad College of Engineering,

Pune.

Abhijit D. Jadhav

Ph.D. CSE Scholar, M.Tech.(CSE),

B.E. (Computer Engineering),

Memberships : IAENG, IAETSD, ACM,

Assistant Professor, Department of Computer Engineering,

Pimpri Chinchwad College of Engineering and Research,

Ravet, Pune.

TECHNICAL PUBLICATIONS

SINCE 1993

An Up Thrust for Knowledge

(i)

DEEP LEARNING

Subject Code : 410251

B.E. (Computer Engineering) Semester : VIII

Published by :

TECHNICAL PUBLICATIONS

An Up Thrust for Knowledge

Amit Residency, Office No.1, 412, Shaniwar Peth,

Pune - 411030, M.S. INDIA Ph. : +91-020-24495496/97

Email : info@technicalpublications.in Website : www.technicalpublications.in

Printer :

Vaishali Printers & Binders

Sr. No. 10/1A,

Ghule Industrial Estate, Nanded Village Road,

Tal. - Haveli, Dist. - Pune - 411041

(ii)

SYLLABUS

Deep Learning - (410251)

Credit	Examination Scheme :
03	In-Sem (Paper) : 30 Marks End-Sem (Paper) : 70 Marks

Unit I Foundations of Deep learning

What is machine learning and deep learning ?,Supervised and Unsupervised Learning, bias variance tradeoff, hyper parameters, under/over fitting regularization, Limitations of machine learning, History of deep learning, Advantage and challenges of deep learning. Learning representations from data, Understanding how deep learning works in three figures, Common Architectural Principles of Deep Network, Architecture Design, Applications of Deep learning, Introduction and use of popular industry tools such as TensorFLow, Keras, PyTorch, Caffe, Shogun. **(Chapter - 1)**

Unit II Deep Neural Networks (DNNs)

Introduction to Neural Networks : The Biological Neuron, The Perceptron, Multilayer Feed-Forward Networks, **Training Neural Networks :** Backpropagation and Forward propagation **Activation Functions :** Linear, Sigmoid, Tannh, Hard Tanh, Softmax, Rectified Linear, **Loss Functions :** Loss Function Notation, Loss Functions for Regression, Loss Functions for Classification, Loss Functions for Reconstruction, **Hyperparameters :** Learning Rate, Regularization, Momentum, Sparsity, Deep Feedforward Networks - Example of Ex OR, Hidden Units, cost functions, error backpropagation, Gradient-Based Learning, Implementing Gradient Descent, vanishing and Exploding gradient descent, Sentiment Analysis, Deep Learning with Pytorch, Jupyter, colab. **(Chapter - 2)**

Unit III Convolution Neural Network (CNN)

Introduction, CNN architecture overview, The Basic Structure of a Convolutional Network-Padding, Strides, Typical Settings, the ReLU layer, Pooling, Fully Connected Layers, The Interleaving between Layers, Local Response Normalization, Training a Convolutional Network. **(Chapter - 3)**

Unit IV Recurrent Neural Networks (RNN)

Recurrent and Recursive Nets : Unfolding Computational Graphs, Recurrent Neural Networks, Bidirectional RNNs, Encoder-Decoder Sequence-to-Sequence Architectures, Deep Recurrent Networks, Recursive Neural Networks, The Challenge of Long-Term Dependencies, Echo State Networks, Leaky Units and Other Strategies for Multiple Time Scales, The Long Short-Term Memory and Other Gated RNNs, Optimization for Long-Term Dependencies, Explicit Memory. **Practical Methodology :** Performance Metrics, Default Baseline Models, Determining Whether to Gather More Data, Selecting Hyper parameters. **(Chapter - 4)**

Unit V Deep Generative Models

Introduction to deep generative model, Boltzmann Machine, Deep Belief Networks, Generative adversarial network (GAN), discriminator network, generator network, types of GAN, Applications of GAN networks. **(Chapter - 5)**

Unit VI Reinforcement Learning

Introduction of deep reinforcement learning, Markov Decision Process, basic framework of reinforcement learning, challenges of reinforcement learning, Dynamic programming algorithms for reinforcement learning, Q Learning and Deep Q-Networks, Deep Q recurrent networks, Simple reinforcement learning for Tic-Tac-Toe. **(Chapter - 6)**

Unit I

1 Foundations of Deep Learning

Syllabus

What is machine learning and deep learning ? Supervised and Unsupervised Learning, bias variance tradeoff, hyper parameters, under/over fitting regularization, Limitations of machine learning, History of deep learning. Advantage and challenges of deep learning. Learning representations from data. Understanding how deep learning works in three figures, Common Architectural Principles of Deep Network, Architecture Design, Applications of Deep learning, Introduction and use of popular industry tools such as TensorFLow, Keras, PyTorch, Caffe, Shogun.

Contents

1.1 Introduction to Machine Learning

- Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) which concerns with developing computational theories of learning and building learning machines.
- **Learning** is a phenomenon and process which has manifestations of various aspects. Learning process includes gaining of new symbolic knowledge and development of cognitive skills through instruction and practice. It is also discovery of new facts and theories through observation and experiment.
- Machine Learning Definition : 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 in T, as measured by P, improves with experience E.
- Machine learning is programming computers to optimize a performance criterion using example data or past experience. Application of machine learning methods to large databases is called **data mining**.
- It is very hard to write programs that solve problems like recognizing a human face. We do not know what program to write because we don't know how our brain does it. Instead of writing a program by hand, it is possible to collect lots of examples that specify the correct output for a given input.
- A machine learning algorithm then takes these examples and produces a program that does the job. The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers. If we do it right, the program works for new cases as well as the ones we trained it on.
- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Another goal is to develop computational models of human learning process and perform computer simulations.
- The goal of machine learning is to build computer systems that can adapt and learn from their experience.
- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instruction. It should carry out to transform the input to output. For example, for addition of four numbers is carried out by giving four number as input to the algorithm and output is sum of all four numbers. For the same task, there may be various algorithms. It is interested to find the most efficient one, requiring the least number of instructions or memory or both.
- For some tasks, however, we do not have an algorithm.

Why Is Machine Learning Important ?

- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine Learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- **Following are some of the reasons :**
  1. Some tasks cannot be defined well, except by examples. For example: recognizing people.
  2. Relationships and correlations can be hidden within large amounts of data. To solve these problems, machine learning and data mining may be able to find these relationships.
  3. Human designers often produce machines that do not work as well as desired in the environments in which they are used.
  4. The amount of knowledge available about certain tasks might be too large for explicit encoding by humans.
  5. Environments change time to time.
  6. New knowledge about tasks is constantly being discovered by humans.
- Machine learning also helps us find solutions of many problems in computer vision, speech recognition and robotics. Machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample.

How Machines Learn ?

- Machine learning typically follows three phases :
  1. **Training :** A training set of examples of correct behavior is analyzed and some representation of the newly learnt knowledge is stored. This is some form of rules.
  2. **Validation :** The rules are checked and, if necessary, additional training is given. Sometimes additional test data are used, but instead, a human expert may validate the rules, or some other automatic knowledge - based component may be used. The role of the tester is often called the opponent.
  3. **Application :** The rules are used in responding to some new situation.

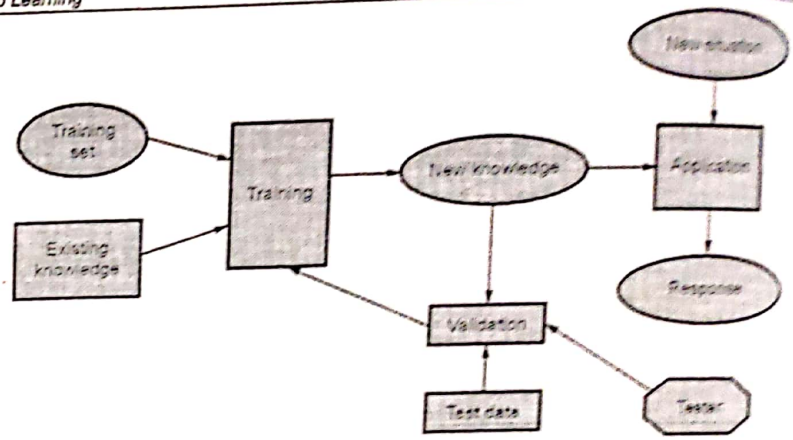


Fig. 1.1.1 Phases of ML

1.1.1 Why Machine Learning is Important ?

- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- **Following are some of the reasons :**
  1. Some tasks cannot be defined well, except by examples. For example Recognizing people.
  2. Relationships and correlations can be hidden within large amounts of data. To solve these problems, machine learning and data mining may be able to find these relationships.
  3. Human designers often produce machines that do not work as well as desired in the environments in which they are used.
  4. The amount of knowledge available about certain tasks might be too large for explicit encoding by humans.
  5. Environments change time to time.
  6. New knowledge about tasks is constantly being discovered by humans.

- Machine learning also helps us find solutions of many problems in computer vision, speech recognition and robotics. Machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample.
- Learning is used when :
  1. Human expertise does not exist (navigating on Mars).
  2. Humans are unable to explain their expertise (speech recognition).
  3. Solution changes in time (routing on a computer network).
  4. Solution needs to be adapted to particular cases (user biometrics).

1.2 What is Deep Learning ?

- The term "deep" usually refers to the number of hidden layers in the neural network.
- Deep learning is a subset of machine learning, which is predicated on idea of learning from example. In machine learning, instead of teaching a computer a massive list of rules to solve the problem, we give it a model with which it can evaluate examples, and a small set of instructions to modify the model when it makes a mistake.
- The basic idea of deep learning is that repeated composition of functions can often reduce the requirements on the number of base functions (computational units) by a factor that is exponentially related to the number of layers in the network.
- Deep learning eliminates some of data pre-processing that is typically involved with machine learning.
- Fig. 1.2.1 shows relation between AI, ML and Deep learning.

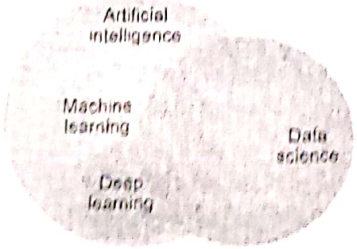


Fig. 1.2.1 Relation between AI, ML and Deep learning

- For example, let's say that we had a set of photos of different pets, and we wanted to categorize by "cat" and "dog". Deep learning algorithms can determine which features (e.g. ears) are most important to distinguish each animal from another. In machine learning, this hierarchy of features is established manually by a human expert.
- In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.
- Deep learning classifies information through layers of neural networks, which have a set of inputs that receive raw data. For example, if a neural network is trained with images of birds, it can be used to recognize images of birds. More layers enable more precise results, such as distinguishing a crow from a raven as compared to distinguishing a crow from a chicken.
- Deep Learning consists of the following methods and their variations :
  a) Unsupervised learning systems such as Boltzman machines for preliminary training, auto-encoders, generative adversarial network.
  b) Supervised learning such as Convolution neural networks which brought technoogy of pattern recognition to a new level.
  c) Recurrent neural networks, allowing to train on processes in time.
  d) Recursive neural networks, allowing to include feedback between circuit elements and chains.

1.2.1 Reasons for Using Deep Learning

1. **Analyzing unstructured data :** Deep learning algorithms can be trained to look at text data by analyzing social media posts, news, and surveys to provide valuable business and customer insights.
2. **Data labelling :** Deep learning requires labeled data for training. Once trained, it can label new data and identify different types of data on its own.
3. **Feature engineering :** A deep learning algorithm can save time because it does not require humans to extract features manually from raw data.
4. **Efficiency :** When a deep learning algorithm is properly trained, it can perform thousands of tasks over and over again, faster than humans.

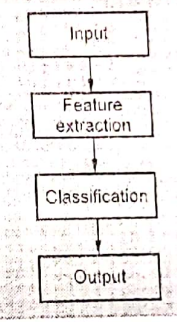
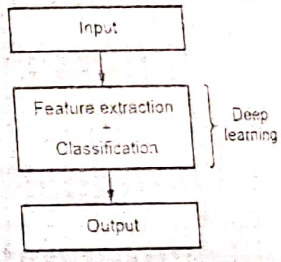
5. **Training :** The neural networks used in deep learning have the ability to be applied to many different data types and applications. Additionally, a deep learning model can adapt by retraining it with new data.

1.2.2 Application of Deep Learning

1. **Aerospace and defense :** Deep learning is utilized extensively to help satellites identify specific objects or areas of interest and classify them as safe or unsafe for soldiers.
2. **Financial services :** Financial institutions regularly use predictive analytics to drive algorithmic trading of stocks, assess business risks for loan approvals, detect fraud, and help manage credit and investment portfolios for clients.
3. **Medical research :** The medical research field uses deep learning extensively. For example, in ongoing cancer research, deep learning is used to detect the presence of cancer cells automatically.
4. **Industrial automation :** The heavy machinery sector is one that requires a large number of safety measures. Deep learning helps with the improvement of worker safety in such environments by detecting any person or objects that comes within the unsafe radius of a heavy machine.
5. **Facial recognition :** This feature utilizing deep learning is being used not just for a range of security purposes but will soon enable purchases at stores. Facial recognition is already being extensively used in airports to enable seamless, paperless check-ins.

1.2.3 Difference between Machine Learning and Deep Learning

Sr. No.	Machine Learning	Deep Learning
1.	Machine learning uses algorithms to parse data, learn from that data, and make informed decisions based on what it has learned.	Deep learning structures algorithms in layers to create an "artificial neural network" that can learn and make intelligent decisions on its own.
2.	Machine learning gives lesser accuracy.	Deep learning gives more accuracy.

3.	Machine learning requires less time for training.	Deep learning requires more time for training.
4.	Needs accurately identified features by human intervention.	It can create new features.
5.	Machine learning models mostly require data in a structured form.	Deep Learning models can work with structured and unstructured data both as they rely on the layers of the Artificial neural network.
6.	Algorithms are detected by data analysts to examine specific variables in data sets.	Algorithms are largely self-directed on data analysis once they are put into production.
7.	Machine learning can work on low-end machines.	Deep learning model needs a huge amount of data to work efficiently, so they need GPU's and hence the high-end machine
8.		

1.2.4 Difference between ML, AI and Data Science

Sr. No.	Machine Learning	Artificial Intelligence	Data Science
1.	Focuses on providing a means for algorithms and systems to learn from experience with data and use that experience to improve over time.	Focuses on giving machines cognitive and intellectual capabilities similar to those of humans.	Focuses on extracting information needles from data haystacks to aid in decision-making and planning.

2.	Machine Learning uses statistical models.	Artificial Intelligence uses logic and decision trees.	Data Science deals with structured data.
3.	A form of analytics in which software programs learn about data and find patterns.	Development of computerized applications that simulate human intelligence and interaction.	The process of using advanced analytics to extract relevant information from data.
4.	Objective is to maximize accuracy.	Objective is to maximize the chance of success.	Objective is to extract actionable insights from the data.
5.	ML can be done through supervised, unsupervised or reinforcement learning approaches.	AI encompasses a collection of intelligence concepts, including elements of perception, planning and prediction.	Uses statistics, mathematics, data wrangling, big data analytics, machine learning and various other methods to answer analytics questions.
6.	ML is concerned with knowledge accumulation.	AI is concerned with knowledge dissemination and conscious machine actions.	Data science is all about data engineering.

1.2.5 Difference between AI, ML and Deep Learning

Sr. No.	AI	ML	DL
1.	AI aims towards building machines that are capable to think like humans.	ML aims to learn through data to solve the problem.	DL aims to build neural networks that automatically discover patterns for feature detection.
2.	AI is subset of data science.	ML is subset of AI and data science.	DL is subset of AI, ML and data science.

3.	All systems of artificial intelligence fall into three types : a) Artificial Narrow Intelligence b) Artificial General Intelligence c) Artificial Super Intelligence	ML algorithms can be broadly classified into three categories a) Supervised learning b) Unsupervised learning c) Reinforcement learning	Deep learning architectures are as follows : a) Convolutional Neural Networks b) Recurrent Neural Networks c) Recursive Neural Networks
4.	Making machines intelligent may or may not need high computational power.	These algorithms can work easily on normal low performance computers without GPUs.	Algorithms are dependent on high performance hardware components that include GPUs.

1.2.6 Advantages and Disadvantages of Deep Learning

Advantages of Deep Learning

- No need for feature engineering.
- DL solves the problem on the end-to-end basis.
- Deep learning gives more accuracy.

Disadvantages of Deep Learning

- DL needs high-performance hardware.
- DL needs much more time to train.
- It is very difficult to assess its performance in real world applications.
- It is very hard to understand.

1.3 Supervised Learning

- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behavior of a system for any set of input values, after an initial training phase.
- **Supervised learning** in which the network is trained by providing it with input and matching output patterns. These input-output pairs are usually provided by an external teacher.

- Human learning is based on the past experiences. A computer does not have experiences.
- A computer system learns from data, which represent some "past experiences" of an application domain.
- To learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved and high-risk or low risk. The task is commonly called : Supervised learning, Classification or inductive learning.
- Training data includes both the input and the desired results. For some examples the correct results (targets) are known and are given in input to the model during the learning process. The construction of a proper training, validation and test set is crucial. These methods are usually fast and accurate.
- Have to be able to generalize : give the correct results when new data are given in input without knowing a priori the target.
- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value.
- A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier or a regression function. Fig. 1.3.1 shows supervised learning process.

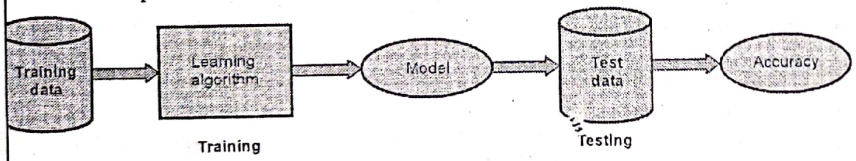


Fig. 1.3.1 Supervised learning process

- The learned model helps the system to perform task better as compared to no learning.
- Each input vector requires a corresponding target vector.

  Training Pair = (Input Vector, Target Vector)
- Fig. 1.3.2 shows input vector. (See Fig. 1.3.2 on next page)

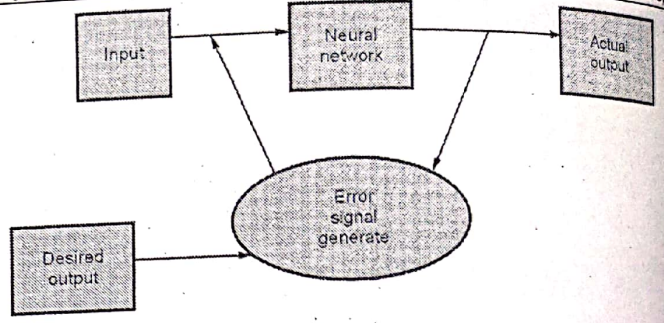


Fig. 1.3.2 Input vector

- Supervised learning denotes a method in which some input vectors are collected and presented to the network. The output computed by the network is observed and the deviation from the expected answer is measured. The weights are corrected according to the magnitude of the error in the way defined by the learning algorithm.
- Supervised learning is further divided into methods which use reinforcement or error correction. The perceptron learning algorithm is an example of supervised learning with reinforcement.

Data formats in supervised learning :

- Supervised learning always uses a dataset to define finite set of real vectors with m features each :

$$X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\} \text{ where } \bar{x}_i \in \Re^m$$

- Considering that user approach is always probabilistic, we need to consider each $\bar{x}_i$ as drawn from a statistical multivariate distribution D. It is also useful to add an important condition upon the whole dataset X. Here we consider that all samples to be independent and identically distributed. This means all variables belong to the same distribution D and considering an arbitrary subset of m values, it happens that :

$$P(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m) = \prod_{i=1}^{m} P(\bar{x}_i)$$

- The corresponding output values can be both numerical - continuous and categorical. In the first case, the process is called regression, while in the second, it is called classification.

- Example : Dataset contains city populations by year for the past 100 years and user want to know what the population of a specific city will be four years from now. The outcome uses labels that already exist in the data set : population, city and year.
- In order to solve a given problem of supervised learning, following steps are performed :
  1. Find out the type of training examples.
  2. Collect a training set.
  3. Determine the input feature representation of the learned function.
  4. Determine the structure of the learned function and corresponding learning algorithm.
  5. Complete the design and then run the learning algorithm on the collected training set.
  6. Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.
- Supervised learning is divided into two types : Classification and Regression.

1. Classification :

- Classification predicts categorical labels (classes), prediction models continuous - valued functions. Classification is considered to be supervised learning.
- Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. Prediction means models continuous-valued functions, i.e., predicts unknown or missing values.
- Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes and data transformation, such as generalizing the data to higher level concepts or normalizing data.
- Numeric prediction is the task of predicting continuos values for given input. For example, we may wish to predict the salary of college employee with 15 years of work experience or the potential sales of a new product given its price.
- Some of the classification methods like back - propagation, support vector machines and k-nearest-neighbor classifiers can be used for prediction.

2. Regression :

- For an input x, if the output is continuous, this is called a regression problem. For example, based on historical information of demand for tooth paste in supermarket, user are asked to predict the demand for the next month.

- Regression is concerned with the prediction of continuous quantities. Linear regression is the oldest and most widely used predictive model in the field of machine learning. The goal is to minimize the sum of the squared errors to fit a straight line to a set of data points.
- Regression algorithm used in supervised learning is linear regression, Bayesian linear regression, polynomial regression, regression tree etc.

1.3.1 Advantages and Disadvantages of Supervised Learning

1. Advantages of supervised learning

- It performs classification and regression tasks.
- It allows estimating or mapping the result to a new sample.
- We have complete control over choosing the number of classes we want in the training data.

2. Disadvantages of supervised learning

- Supervised learning cannot handle all complex tasks in Machine Learning.
- Computation time is vast for supervised learning.
- It requires a labelled data set.
- It requires a training process.

1.4 Unsupervised Learning

- Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.
- In unsupervised learning, a dataset is provided without labels and a model learns useful properties of the structure of the dataset. The main goal of unsupervised learning is to discover hidden and interesting patterns in unlabeled data.
- They are called unsupervised because they do not need a teacher or super-visor to label a set of training examples. Only the original data is required to start the analysis.
- Unsupervised learning tasks typically involve grouping similar examples together, dimensionality reduction and density estimation.
- Common algorithms used in unsupervised learning include clustering, anomaly detection, neural networks.

- Fig. 1.4.1 shows unsupervised learning.

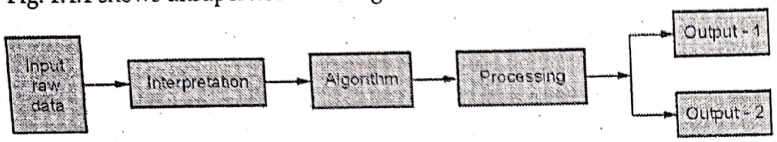


Fig. 1.4.1 Unsupervised learning

- The most common unsupervised learning method is cluster analysis, which applies clustering methods to explore data and find hidden patterns or groupings in data. Unsupervised learning is typically applied before supervised learning, to identify features in exploratory data analysis and establish classes based on groupings.
- Unsupervised machine learning is mainly used to :
  1. Cluster datasets on similarities between features or segment data.
  2. Understand relationship between different data point such as automated music recommendations.
  3. Perform initial data analysis.
- Unsupervised learning algorithms have the capability of analyzing large amounts of data and identifying unusual points among the dataset. Once those anomalies have been detected, they can be brought to the awareness of the user, who can then decide to act or not on this warning.
- Anomaly detection can be very useful in the financial and banking sectors. Indeed, financial fraud has become a daily problem, due to the ease with which credit card details can be stolen. Using unsupervised learning models, unauthorized or fraudulent transactions on a bank account can be identified as it will most often constitute a change in the user's normal pattern of spending.
- Example : Using customer data and user want to create segments of customers who like similar products. The data that user are providing is not labeled and the labels in the outcome are generated based on the similarities that were discovered between data points.
- Types of unsupervised learning is clustering and association analysis.
- There is a wide range of algorithms that can be deployed under unsupervised learning. A few of them includes : K-means clustering, Principal component analysis, Hierarchical clustering and Dendrogram.

1.4.1 Advantages and Disadvantages of Unsupervised Learning

1. Advantages of unsupervised learning

- It does not require a training data to be labelled.
- Dimensionality reduction can be easily accomplished using unsupervised learning.
- Capable of finding previously unknown patterns in data.

2. Disadvantages of unsupervised learning

- Difficult to measure accuracy or effectiveness due to lack of predefined answers during training.
- The results often have lesser accuracy.
- The user needs to spend time interpreting and label the classes which follow that classification.

1.4.2 Difference between Supervised and Unsupervised Learning

Sr. No.	Supervised learning	Unsupervised learning
1.	Desired output is given.	Desired output is not given.
2.	It is not possible to learn larger and more complex models than with supervised learning.	It is possible to learn larger and more complex models with unsupervised learning.
3.	Use training data to infer model.	No training data is used.
4.	Every input pattern that is used to train the network is associated with an output pattern.	The target output is not presented to the network.
5.	Trying to predict a function from labeled data.	Try to detect interesting relations in data.
6.	Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given.	For unsupervised learning typically either the target variable is unknown or has only been recorded for too small a number of cases.
7.	Example : Optical character recognition.	Example : Find a face in an image.

8.	We can test our model.	We can not test our model.
9.	Supervised learning is also called classification.	Unsupervised learning is also called clustering.

1.5 Bias Variance Tradeoff

- In the experimental practice we observe an important phenomenon called the bias variance dilemma.
- In supervised learning, the class value assigned by the learning model built based on the training data may differ from the actual class value. This error in learning can be of two types, errors due to 'bias' and error due to 'variance'.
- Fig. 1.5.1 shows bias-variance trade off.

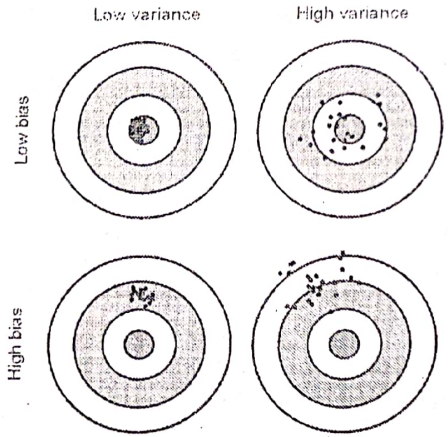


Fig. 1.5.1 Bias-variance trade off

- Give two classes of hypothesis (e.g. linear models and k-NNs) to fit to some training data set, we observe that the more flexible hypothesis class has a low bias term but a higher variance term. If we have parametric family of hypothesis, then we can increases the flexibility of the hypothesis but we still observe the increase of variance.
- The bias-variance-dilemma is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithm from generalizing beyond their training set :

1. The bias is error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs.
2. The variance is error from sensitivity to small fluctuations in the training set. High variance can cause overfitting : modeling the random noise in the training data, rather than the intended outputs.

- In order to reduce the model error, the designer can aim at reducing either the bias or the variance, as the noise components is irreducible.
- As the model increases in complexity, its bias is likely to diminish. However, as the number of training examples is kept fixed, the parametric identification of the model may strongly vary from one DN to another. This will increase the variance term.
- At one stage, the decrease in bias will be inferior to the increase in variance, warning that the model should not be too complex. Conversely, to decrease the variance term, the designer has to simplify its model so that it is less sensitive to a specific training set. This simplification will lead to a higher bias.

Example 1.5.1

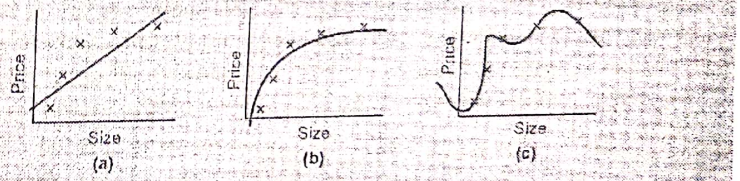


Fig. 1.5.2

Explain the above Fig. 1.5.2 (a), (b) and (c).

Solution :

- Given Fig. 1.5.2 is related to overfitting and underfitting.

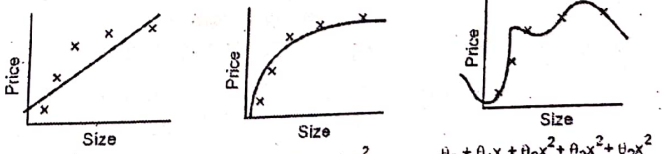


Fig. 1.5.3

Underfitting (High bias and low variance) :

- A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.
- It usually happens when we have less data to build an accurate model and also when we try to build a linear model with a non-linear data.
- In such cases the rules of the machine learning model are too easy and flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions.
- Underfitting can be avoided by using more data and also reducing the features by feature selection.

Overfitting (High variance and low bias) :

- A statistical model is said to be overfitted, when we train it with a lot of data.
- When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set.
- Then the model does not categorize the data correctly because of too many details and noise.
- The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.
- A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

1.6 Generalization : Overfitting vs Underfitting

- In addition to using models for prediction, the ability to interpret what a model has learned is receiving an increasing amount of attention.
- Interpretability has to do with how accurate a machine learning model can associate a cause to an effect.
- If a model can take the inputs, and routinely get the same outputs, the model is interpretable :
  1. If you overeat your magi at dinnertime and you always have troubles sleeping, the situation is interpretable.
  2. If all 2019 polls showed " ABC party" win and the "XYZ party" candidate took office, all those models showed low interpretability.
- Interpretability poses no issue in low-risk scenarios. If a model is recommending movies to watch, that can be a low-risk task

- Fitness of a target function approximated by a learning algorithm determines how correctly it is able to classify a set of data it has never seen.

1.6.1 Underfitting and Overfitting

- Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to overfitting and poor generalization.
- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Overfitting is when a classifier fits the training data too tightly. Such a classifier works well on the training data but not on independent test data. It is a general problem that plagues all machine learning methods.
- Underfitting : If we put too few variables in the model, leaving out variables that could help explain the response, we are underfitting. Consequences :
  1. Fitted model is not good for prediction of new data - prediction is biased
  2. Regression coefficients are biased
  3. Estimate of error variance is too large
- Because of overfitting, low error on training data and high error on test data. Overfitting occurs when a model begins to memorize training data rather than learning to generalize from trend.
- The more difficult a criterion is to predict, the more noise exists in past information that need to be ignored. The problem is determining which part to ignore.
- Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. We can determine whether a predictive model is underfitting or overfitting the training data by looking at the prediction error on the training data and the evaluation data. Fig. 1.6.1 shows underfitting and overfitting.

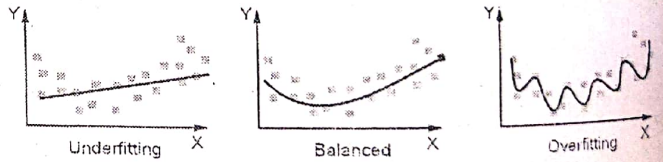


Fig. 1.6.1

- Reasons for overfitting
  1. Noisy data
  2. Training set is too small
  3. Large number of features

- In the machine learning the more complex model is said to show signs of overfitting, while the simpler model underfitting. Often several heuristic are developed in order to avoid overfitting, for example, when designing neural networks one may :
  1. Limit the number of hidden nodes
  2. Stop training early to avoid a perfect explanation of the training set, and
  3. Apply weight decay to limit the size of the weights, and thus of the function class implemented by the network

1.7 History of Deep Learning

- The history of Deep Learning can be traced back to 1943, when Walter Pitts and Warren McCulloch created a computer model based on the neural networks of the human brain. They used a combination of algorithms and mathematics they called "threshold logic" to mimic the thought process
- There basic aim was to mimic thought process of human brain; they used algorithms and mathematics to make the threshold logic to mimic human thought process. Alan Turing called the father of AI concluded in 1951 that the machines would not take much time in started thinking of their own; at some point of time, they would be able to talk to each other and it is also expected that they would take the control of the universe.
- Warren McCulloch and Walter Pitts used a combination of mathematics and algorithms they called threshold logic to mimic the thought process. Since then, deep learning has evolved steadily, over the years with two significant breaks in its development.
- The development of the basics of a continuous Back Propagation Model is credited to Henry J. Kelley in 1960. Stuart Dreyfus came up with a simpler version based only on the chain rule in 1962. The concept of back propagation existed in the early 1960s but only became useful until 1985.
- The next significant evolutionary step for deep learning took place in 1999, when computers started becoming faster at processing data and graphics processing units were developed. Neural networks also have the advantage of continuing to improve as more training data is added.
- Around the year 2000, The Vanishing Gradient Problem appeared. It was discovered "features" formed in lower layers were not being learned by the upper layers, because no learning signal reached these layers.

- In 2001, a research report by META Group described the challenges and opportunities of data growth as three-dimensional. The report described the increasing volume of data and the increasing speed of data as increasing the range of data sources and types. This was a call to prepare for the onslaught of Big Data, which was just starting.
- In 2009, Fei-Fei Li, an AI professor at Stanford launched ImageNet, assembled a free database of more than 14 million labeled images. The Internet is and was, full of unlabeled images. Labeled images were needed to "train" neural nets.
- By 2011, the speed of GPUs had increased significantly, making it possible to train convolutional neural networks "without" the layer-by-layer pre-training. With the increased computing speed, it became obvious deep learning had significant advantages in terms of efficiency and speed.
- Generative Adversarial Network (GAN) is a class of machine learning system invented by Ian Goodfellow and his colleagues in 2014. Coming up in history in 2016 Google DeepMind challenge match between Alpha Go versus Lee Sedol, the AlphaGo win all the matches from a world champion Lee Sedol.
- AlfaGo and AlfaZero are computer programs developed by artificial intelligence research company called DeepMind in (2016 - 2017); it plays the board game Go.
- The transformer introduced in 2017 - 19 a deep learning model used specially used for Natural language Processing (NLP).
- Although there is a lot of community contributed to the deep learning but Yann LeCun, Geoffrey Hinton, and Yoshua Bengio have received Turing awards in 2018.

1.8 Learning representations from Data

- A machine learning model can't directly see, hear, or sense input examples. Instead, we must create a representation of the data to provide the model with a useful vantage point into the data's key qualities. That is, in order to train a model, we must choose the set of features that best represent the data.
- Representation learning is a class of machine learning approaches that allow a system to discover the representations required for feature detection or classification from raw data. The requirement for manual feature engineering is reduced by allowing a machine to learn the features and apply them to a given activity.

- In representation learning, data is sent into the machine and it learns the representation on its own. It is a way of determining a data representation of the features, the distance function, and the similarity function that determines how the predictive model will perform.
- Representation learning works by reducing high-dimensional data to low-dimensional data, making it easier to discover patterns and anomalies while also providing a better understanding of the data's overall behavior.
- Representation learning is concerned with training machine learning algorithms to learn useful representations.
- Deep neural networks can be considered representation learning models that typically encode information which is projected into a different subspace. These representations are then usually passed on to a linear classifier to, for instance, train a classifier.
- Representation learning can be divided into :
  a) Supervised representation learning : learning representations on task A using annotated data and used to solve task B
  b) Unsupervised representation learning : learning representations on a task in an unsupervised way. These are then used to address downstream tasks and reducing the need for annotated data when learning news tasks.

1.8.1 Greedy Layer - Wise Unsupervised Pretraining

- Greedy algorithms break a problem into many components, then solve for the optimal version of each component in isolation. Unfortunately, combining the individually optimal components is not guaranteed to yield an optimal complete solution.
- Greedy Layer-Wise Unsupervised Pretraining relies on single-layer representation learning algorithm. Each layer is pretrained using unsupervised learning, taking the output of previous layer and producing as output a new representation of the data, whose distribution is hopefully simpler.

Why it is called Greedy layer-wise pretraining ?

- Greedy because, it is a greedy algorithm that optimizes each piece of the solution independently
- Layer-wise because, independent pieces are the layers of the network and training proceeds one layer at a time.
- Pretraining because, it is only a first step before applying a joint training algorithm is applied to fine-tune all layers together.

- Unsupervised feature learning algorithm L, which takes a training set of examples and returns an encoder or feature function f. The X is raw input data.

$f \leftarrow$ Identity function

$\bar{X} = X$

for k = 1, ..., m do

$f^{(k)} = \mathcal{L}(\bar{X})$

$f \leftarrow f^{(k)} \circ f$

$\bar{X} \leftarrow f^{(k)}(\bar{X})$

end for

if fine-tuning then

$f \leftarrow T(f.X.Y)$

end if

Return f

- Gready : Optimize each piece of the solution independently, on piece at a time.
- Layer-Wise : The independent pieces are the layer of the network. Training proceeds once layer at a time, training the k^{th} layer while keeping the previous ones fixed.
- Unsupervised : each layer is trained with an unsupervised representation learning algorithm

When and why does unsupervised pretraining ideas work ?

1. **Idea1 :** Choice of initial parameters for a deep NN can have a significant regularizing effect on the model.

- It remains possible that pretraining initializes the model in a location that would otherwise be inaccessible. For example : a region surrounded by areas where the cost function varies so much from one example to another that mini-batches give only a very noisy estimate of the gradient a region surrounded by areas where the Hessian matrix is so pooly conditioned that gradient descent methods must use very small steps.
- We cannot characterize the exactly what aspects of the pretrained parameters are retained during the supervised training stages.

2. **Idea2 :** Learning about the input distribution can help with learning about the mapping from input to output.

- Some features that are useful for the unsupervised task may also be useful for supervised task. This is not yet understood at a mathematical, theoretical level. Many aspects of this approach are highly dependent on the specific models used.
- For example, if we wish to add a linear classifier on top of pretrained features, the features mush make the underlying classes linearly separable. This is another reason that simultaneous supervised and unsupervised learning can be preferable.

Disadvantages of Unsupervised Pretraining :

1. Unsupervsied pretraining does not offer a clear way to adjust the strength of the regularization arising from the unsupervised stage. When we perform unsupervised and supervised learning simultaneously, instead of using the pretraining strategy, there is a single hyperparameter, usually a coefficient attatched to the unsupervised cost, that determine hows strongly supervsied objective will regularize the supervsied model.
2. Two separate training phases has its own hyperparameters. The performance of the second phase cannot be predicted during the first phase, so there is a long delay between proposing hyperparameters for the first phase and being able to update them using feedback from the second phase. Most principled approach: use validation set error in the supervised phase to select hyperparameters of the pretraining phase

1.8.2 Transfer Learning and Domain Adaption

- Transfer learning and domain adaptation refer to situation where what has been learned in one setting is exploited to improve generalization in another settings.
- Transfer learning, multitask learning and domain adaptation can be achieved via representation learning when there exist features that are useful for different settings or tasks, corresponding to underlying factors that appear in more than one setting.
- Two extreme form of transfer learning :
  1. **One-shot learning :** Only one example of transfer task is given for one-shot learning. It is possible because the representation learns cleanly separate the underlying classes during first stage. During the transfer learning stage, only one labeled example is needed to infer the label of many possible test examples that all cluster around the same point in representation space.

2. **Zero-shot learning :** No labeled examples are given at all. It is only possible because additional information has been exploited during training.

1.9 Understanding How Deep Learning Works in Three Figures

- In machine learning we take some data, train a model on that data, and use the trained model to make predictions on new data. The process of training a model can be seen as a learning process where the model is exposed to new, unfamiliar data step by step.
- At each step, the model makes predictions and gets feedback about how accurate its generated predictions were. This feedback, which is provided in terms of an error according to some measure is used to correct the errors made in prediction.
- Deep learning algorithms attempt to draw similar conclusions as humans would by continually analyzing data with a given logical structure. To achieve this, deep learning uses a multi-layered structure of algorithms called neural networks.
- Neural networks enable us to perform many tasks, such as clustering, classification or regression. With neural networks, we can group or sort unlabeled data according to similarities among samples in the data.
- Deep Learning is a machine learning method. It allows us to train an AI to predict outputs, given a set of inputs. Both supervised and unsupervised learning can be used to train the AI.
- Let us consider airplane ticket price estimation service. We want our airplane ticket price estimator to predict the price using the following inputs :
  1. Origin Airport
  2. Destination Airport
  3. Departure Date
  4. Airline
- The neurons are grouped into three different types of layers : Input Layer, Hidden Layer(s) and Output Layer.
- The input layer receives input data. Here we have four neurons in the input layer: Origin city, Destination city, Departure Date, and Travel Company. The input layer passes the inputs to the first hidden layer.
- The hidden layers perform mathematical computations on inputs. One of the challenges in creating neural networks is deciding the number of hidden layers, as well as the number of neurons for each layer.

- The "Deep" in Deep Learning refers to having more than one hidden layer. The output layer returns the output data. In this case, it gives us the price prediction. Fig. 1.9.1 shows deep neural network.

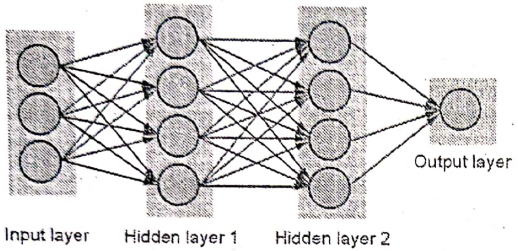


Fig. 1.9.1 Deep neural network

- Fig. 1.9.2 shows ticket reservation using deep neural network.

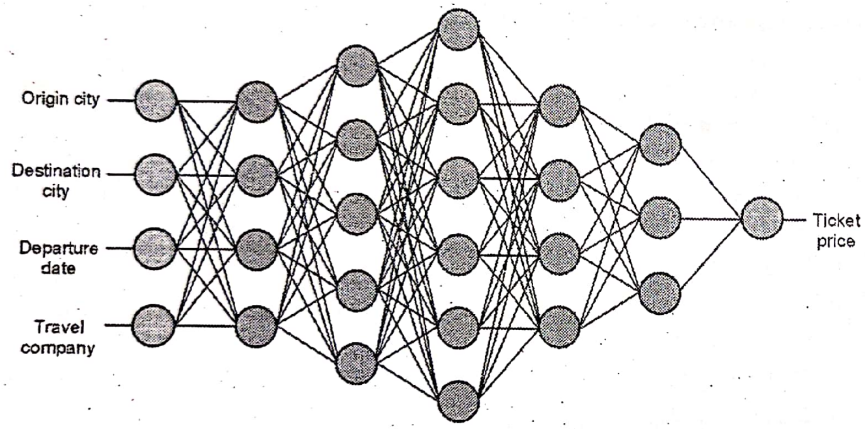


Fig. 1.9.2 Ticket reservation using deep neural network

- Each connection between neurons is associated with a weight. This weight dictates the importance of the input value. The initial weights are set randomly. When predicting the price of an travel ticket, the departure date is one of the heavier factors. Hence, the departure date neuron connections will have a big weight.
- Each neuron has an Activation Function. One of its purposes is to "standardize" the output from the neuron. Once a set of input data has passed through all the layers of the neural network, it returns the output data through the output layer.
- Training phase in deep network is complex part because it requires large data set and a large amount of computational power.

- In this example, we need to find historical data of ticket prices. And due to the large amount of possible city and departure date combinations, we need a very large list of ticket prices. To train the AI, we need to give it the inputs from our data set and compare its outputs with the outputs from the data set. Since the AI is still untrained, its outputs will be wrong.
- Once we go through the whole data set, we can create a function that shows us how wrong the AI's outputs were from the real outputs. This function is called the Cost Function. Ideally, cost function to be zero. Gradient Descent technique used for reducing cost. Fig. 1.9.3 shows weight with cost function.

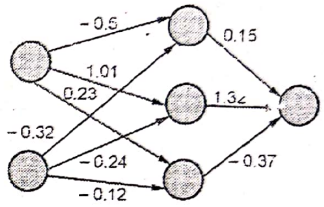


Fig. 1.9.3 Weight with cost function

- The weights of the network are assigned random values at the start. Therefore the network only implements a series of random transformations. Its output is away from what it should ideally be. The loss score is accordingly very high. The loss score decreases as the weights are adjusted a little in the correct direction. This training loop repeated a sufficient number of times.. This yields weight values that minimize the loss function.
- It works by changing the weights in small increments after each data set iteration. By computing the derivative (or gradient) of the cost function at a certain set of weight. Updating the weights using gradient descent is done automatically. This is working of deep network.

1.10 Common Architectural Principles of Deep Network

- The purposes of understanding deep networks are as follows :
  a) Parameters b) Layers c) Activation functions
  d) Loss functions e) Optimization methods f) Hyperparameters
- Parameters : Parameters in neural networks relate directly to the weights on the connections in the network. Gradient descent method is used to find good values for the parameter vector to minimize loss across training dataset.

- **Layers :** Layers are a fundamental architectural unit in deep networks. Sometime we need to customize a layer by changing the type of activation function.
- **Activation Functions :** To drive feature extraction, activation functions are used in specific architectures. The higher-order features learned from the data in deep networks are a nonlinear transform applied to the output of the previous layer. This allows the network to learn patterns in the data within a constrained space.
- **Loss Functions :** Loss functions quantify the agreement between the predicted output and the actual output. Various types of loss functions are squared loss, logistic loss, hinge loss and negative log likelihood.
- **Optimization Algorithms :** Training a model in machine learning involves finding the best set of values for the parameter vector of the model. Optimization algorithms are of two types : First-order and Second-order.
- First-order optimization algorithms calculate the Jacobian matrix. Second-order algorithms calculate the derivative of the Jacobian by approximating the Hessian. Second-order methods take into account interdependencies between parameters when choosing how much to modify each parameter.
- **Hyper - parameters :** Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. Various categories of hyper-parameters are as follows :
  a) Layer size
  b) Magnitude (momentum, learning rate)
  c) Regularization (dropout, drop connect, L1, L2)
  d) Activations (and activation function families)
  e) Weight initialization strategy
  f) Loss functions
  g) Settings for epochs during training (mini-batch size)
  h) Normalization scheme for input data (vectorization)

1.10.1 Optimization Algorithms

- Optimization issue in machine learning algorithms is about finding the correct set of inputs for a function that gives the maximum effectiveness in the function evaluation.

- Be it fitting in the logical regression models in machine learning or training the neural networks with varying datasets, the problem of optimization arises in all phases of a machine learning model.
- Out of hundreds of optimization algorithms available, it becomes difficult to choose a single algorithm that can give the best performance while applied in our machine learning model. One suggestable approach to execute these optimization algorithms and make the maximum use of it is to group these algorithms and execute them on the machine learning.
- Most of the times, the "continuous function optimization" problem arises in majority of the machine learning algorithms, in which most of the input given are the real numbers, and so are the outputs. These kinds of problems that take only discrete values as input are generally referred to as "Combinatorial Optimization Problems".
- The optimization algorithms state that if more information about the target function can be made available, it becomes easier to optimize that function and that information can also be utilized effectively for further data processing.
- The major point that comes across the execution of an optimization algorithm is to decide whether we can differentiate an objective function at a given point or not.
- That is, can we calculate the first derivative of a function for a given solution or not ?
- Based on this point, the optimization algorithms are further classified into two categories : One that differentiates the function and other that does not.
- Hence, in this section, we shall discuss the "differentiable" and "non differentiable" objective functions that can be used to group multiple optimization algorithms.

1.10.2 Differentiable Objective Function

- If we can calculate the derivative of a function at any given point while input is given to the system such function can be referred to as "Differential Function".
- One can define the derivative of a function as the rate at which the function changes its value at a given point of time. This is often referred to as a slope too.
- One can apply the optimization techniques on these derivative functions using simple calculus.
- Optimization techniques can sound easier if the derivatives of these "continuous functions", as mentioned above, can be calculated. Some of the algorithms that uses these gradient values of the derivatives are as follows :

1. **Bracketing algorithms :**
   - This technique is helpful when there are problems having only one input variable and the optima exist in the pre - defined specific criteria or range.
   - These algorithms can easily navigate this range that is known already and locate the optima. The only drawback is that the algorithm assumes that there is only one optima present in the model.
   - The advantage of using this algorithm is that it can be utilized in a model even if sometimes there is no derivative information about the variables available.
   - Some of the examples that use bracketing algorithms are Fibonacci search, Golden Section search, Bisection method, etc.
2. **Local descent algorithms :**
   - These algorithms work for the models in which there are multiple input variables with one global optima.
   - The algorithm is widely used in the line search problem.
   - This problem includes the definition of the direction to move during a search space and then it performs the bracketing type search in a line in the direction chosen.
   - The algorithm executes until no other iteration of finding the improved directions is possible.
   - These iterations make the algorithm costly as it continues its execution till an effective direction is obtained.
3. **First order algorithms :**
   - These algorithms use the first order derivatives (gradient) to decide the direction to move in the search space.
   - This algorithm works by first calculating the first derivative of the function, and then following it in the opposite direction, for example going downhill to minimum value for minimization problems, with the help of step size, also known as "learning rate".
   - This step size, or learning rate, is a hyperparameter in the algorithm, that decides the distance to cover, or how far to cover in a search space, which is opposite to the normally used local descent algorithms, which do not have this hyperparameter and performs a full line search in every directions specified.

   - These algorithms are also known as "Gradient Descent" algorithms and following the introduction to some of the minor extensions, these are also known as Momentum, Adagrad, RMSProp, Adam, etc.
   - These gradient descent algorithms are also helpful in training the artificial neural networks and implementing deep learning models in it, by providing the template for Stochastic Gradient Descent, useful for artificial neural networks.
   - Here, the gradient will be based on assumption, instead of direct calculation, using the prediction techniques on the trained data.
4. **Second order algorithms :**
   - These algorithms use the second order derivative of the input variables for choosing the direction of movement in the search space.
   - The algorithms work appropriately only for the objective functions in which the Hessian matrix needs to be calculated.
   - Some of the examples in which the second order algorithms are used.
     - Newton's method
     - Secant method
   - These algorithms are also known as Quasi Newton Methods.

1.10.3 Non - Differentiable Objective Function

- Although, optimization algorithms working on the derivatives of the objective functions are efficient and fast, there are certain objective functions whose derivatives cannot be calculated, the reason being the complexity of the function.
- Some of the reasons for the complexities in the function include.
  - Lack of analysis of the function
  - Multiple optima required.
  - Evaluation of stochastic functions
  - Objective functions are discontinuous.
- The optimization algorithms that do not make the compulsion for the first or second order derivatives for their objective functions are called as Black - box optimization algorithms. Some of these algorithms are :
  - Direct algorithms
  - Stochastic algorithms
  - Population algorithms

Let us have a brief of each of these :

1. **Direct algorithms :**
   - These algorithms are used when the calculation of the derivatives of the objective function is not possible.
   - The algorithms work with an assumption that the objective function contains single optima.
   - These methods are also referred to as "pattern search" algorithms, since they analyze the search space using the geometrical shapes and patterns.
   - The gradient information required to run the algorithm is calculated directly from the objective function by computing the difference between the scores obtained from the points in the search space.
   - This information estimated are then helpful in choosing a direction to commute in the search space and cover the region of the present optima.
   - Some of the examples that use these direct algorithms are Cyclic Coordinate search, Powell's method, Hooke-Jeeves method, etc.
2. **Stochastic algorithms :**
   - For the variables whose derivatives cannot be calculated, stochastic optimization algorithms use the randomness for those objective functions to commute in the search space.
   - Hence, due to the randomness involved, the stochastic algorithms involve many data sampling for the objective function.
3. **Population algorithms :**
   - These algorithms maintain a pool of solutions for a given input, often known as a population of candidate solutions, which are used to explore, sample the optima.
   - These algorithms are mostly used in the problems that are more challenging and also involves the evaluation of functions containing considerable noise in it, in addition to the presence of multiple global optima. The solutions of such algorithms are difficult to be found by other methods.
   - Genetic algorithms, differential evolution, particle swarm optimization, etc. are the examples of population algorithms.

1.11 Architecture Design

- Architecture refers to the overall structure of the network.
- Most neural networks are organized into groups of units called layers. These layers

are arranged in a chain structure, with each layer being a function of the layer that preceded it.

- In this structure, the first layer is given by

$$h^{(1)} = g^{(1)}(W^{(1)T}x + b^{(1)})$$

The second layer is given by

$$h^{(2)} = g^{(2)}(W^{(2)T}x + b^{(2)})$$

1.12 Applications of Deep Learning

1. **Aerospace and defense :** Deep learning is utilized extensively to help satellites identify specific objects or areas of interest and classify them as safe or unsafe for soldiers.
2. **Financial services :** Financial institutions regularly use predictive analytics to drive algorithmic trading of stocks, assess business risks for loan approvals, detect fraud and help manage credit and investment portfolios for clients.
3. **Medical research :** The medical research field uses deep learning extensively. For example, in ongoing cancer research, deep learning is used to detect the presence of cancer cells automatically.
4. **Industrial automation :** The heavy machinery sector is one that requires a large number of safety measures. Deep learning helps with the improvement of worker safety in such environments by detecting any person or objects that comes within the unsafe radius of a heavy machine.
5. **Facial recognition :** This feature utilizing deep learning is being used not just for a range of security purposes but will soon enable purchases at stores. Facial recognition is already being extensively used in airports to enable seamless, paperless check-ins.

1.12.1 Speech Recognition Problem

- Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.
- Automatic Speech Recognition (ASR) is the task of producing a text transcription of the audio signal from a speaker.
- Large Vocabulary Speech Recognition (LVSR) involves a large or open vocabulary and we will also take it to imply a speaker independent recognizer.

- The most general LVSR systems may also need to operate on spontaneous as well as read speech with audio data from an uncontrolled recording environment corrupted by both unstructured and structured noise
- Traditional LVSR recognizers are based on Hidden Markov Models (HMMs) with Gaussian Mixture Model (GMM) emission distributions, n-gram language models, and use beam search for decoding.
- GMM-HMM acoustic models are trained with the Expectation-Maximization (EM) algorithm using a procedure that trains progressively more sophisticated models using an initial alignment from the previous model.
- In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal.
- The feature extraction is usually performed in three stages.
  1. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectro temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals.
  2. The second stage compiles an extended feature vector composed of static and dynamic features.
  3. Finally, the last stage transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer.
- In speech recognition a supervised pattern classification system is trained with labeled examples; that is, each input pattern has a class label associated with it.
- Pattern classifiers can also be trained in an unsupervised fashion. For example in a technique known as vector quantization, some representation of the input data is clustered by finding implicit groupings in the data.
- The resulting table of cluster centers is known as a codebook, which can be used to index new vectors by finding the cluster center that is closest to the new vectors.
- Once a feature selection or classification procedure finds a proper representation, a classifier can be designed using a number of possible approaches

1.12.2 Image Classification

- Image classification forms basis of computer vision problems.
  1. In self driving cars image classification can be used to detect traffic lights, trees, vehicles, peoples around etc.

  2. In healthcare it can be used to analyze medical images and depict the symptoms of illness.
  3. With the ubiquitous technologies such as AI and IOT huge amount of data in the form of images, video, speech is generated. Image and video data posted by persons can be used in recommendation system in online shopping or places to visit etc.
- Image classification is where a computer can analyses an image and identify the 'class' the image falls under. For example, input an image of a sheep. Image classification is the process of the computer analyzing the image and telling you it's a sheep.
- Early image classification relied on raw pixel data. This meant that computers would break down images into individual pixels. The problem is that two pictures of the same thing can look very different. They can have different backgrounds, angles, poses, etc. This made it quite the challenge for computers to correctly 'see' and categories images.
- Image classification with deep learning most often involves convolutional neural networks or CNNs. In CNNs, the nodes in the hidden layers don't always share their output with every node in the next layer.
- Deep learning allows machines to identify and extract features from images. This means they can learn the features to look for in images by analyzing lots of pictures.

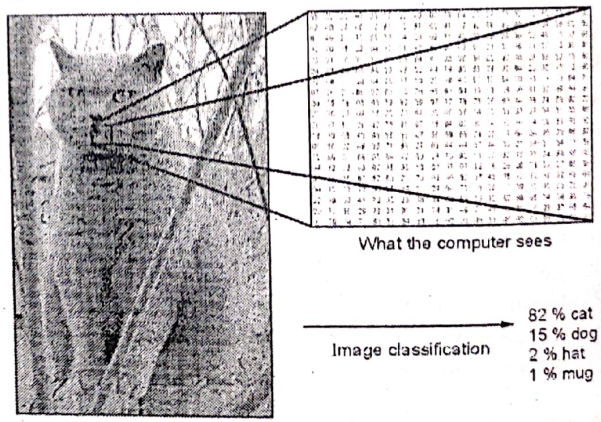


Fig. 1.12.1

- Image is analyzed in the form of pixels. Image is an array of pixels where size of the

matrix depends on resolution of an image. Image classification task is done by grouping pixels into specified categories referred to as classes.

- The image is segregated into its most prominent features using the algorithm giving the classifier an idea about the class of the image it may belong. Thus the feature extraction process is most important step in image classification. Also data fed to the algorithm plays important role particularly in supervised image classification technique.

1.13 Introduction and use of Popular Industry Tools

1.13.1 TensorFLow

- TensorFlow is an end-to-end open source platform for Deep learning. It was developed by the Google Brain team and released in November 2015.
- It uses Python or JavaScript to provide a convenient front-end API for building applications, while executing those applications in high-performance C++.
- TensorFlow is popular for its TensorFlow Core programs where it has two main actions.
  1. Building the computational graph in the construction phase
  2. Running the computational graph in the execution phase
- A tensor is a mathematical object and a generalization of scalars, vectors, and matrices. A tensor can be represented as a multidimensional array. Let's understand how TensorFlow works.
  1. Its programs are usually structured into a construction phase and an execution phase.
  2. The construction phase assembles a graph that has nodes (ops/operations) and edges (tensors).
  3. The execution phase uses a session to execute ops (operations) in the graph.
- Features of Tensorflow are as follows :
  1 Faster debugging with Python tools
  2 Dynamic models with Python control flow
  3 Support for custom and higher-order gradients
  4 TensorFlow offers multiple levels of abstraction, which helps to build and train models.
  5 TensorFlow provides the flexibility and control with features like the Keras functional API and model.

  6 Well-documented so easy to understand.
  7 Probably the most popular easy to use with Python.
- TensorFlow programs use a tensor data structure to represent all data, only tensors are passed between operations in the computation graph. TensorFlow allows developers to create dataflow graphs, structures that describe how data moves through a graph or a series of processing nodes.
- Each node in the graph represents a mathematical operation, and each connection or edge between nodes is a multidimensional data array, or tensor.
- TensorFlow applications can be run on most any target that's convenient : a local machine, a cluster in the cloud, iOS and Android devices, CPUs or GPUs. If you use Google's own cloud, you can run TensorFlow on Google's custom TensorFlow Processing Unit (TPU) silicon for further acceleration. The resulting models created by TensorFlow, though, can be deployed on most any device where they will be used to serve predictions.
- Tesorflow supports three main type of data types : Constants, Variables and Placeholders.
- Constant in TensorFlow : tf.constant() function is used to create constant in TensorFlow. Here, it will always create host/CPU tensor. Syntax is as follows:

  tf.constant (value, dtype=None, shape=Name, name='Constant')

  where

  a) **Value :** A constant value or list. This value is the constant value that is passed to the function.

  b) **Dtype :** It is the data type of the element. dtype is passed to the specify the datatype of the value which is passed as constant. If dtype is not mentioned, then it concludes the type from value, by default.

  c) **Shape :** It is optional dimensions given to the value.

  d) **Name :** Another optional attribute which assigns name for the tensor, which contains the constant value.

  e) **Verify_shape :** It is a boolean value that helps enabling of verification of the shape of values

1.13.2 Keras

- Keras is a high-level neural networks API, capable of running on top of Tensorflow, Theano, and CNTK. It enables fast experimentation through a high level,

userfriendly, modular and extensible API. Keras can also be run on both CPU and GPU.

- Keras is written in Python and supports multiple back-end neural network computation engines.
- Keras provides seven different datasets, which can be loaded in using Keras directly. These include image datasets as well as a house price and a movie review datasets.
- Salient features of Keras
  1. Keras is a high-level interface and uses Theano or Tensorflow for its backend.
  2. It runs smoothly on both CPU and GPU.
  3. Keras supports almost all the models of a neural network – fully connected, convolutional, pooling, recurrent, embedding, etc. Furthermore, these models can be combined to build more complex models.
  4. Keras, being modular in nature, is incredibly expressive, flexible, and apt for innovative research.
  5. Keras is a completely Python-based framework, which makes it easy to debug and explore.
- The model is the core Keras data structure. There are two main types of models available in Keras: the Sequential model, and the Model class used with the functional API.
- The easiest way of creating a model in Keras is by using the sequential API, which lets we stack one layer after the other. The problem with the sequential API is that it doesn't allow models to have multiple inputs or outputs, which are needed for some problems.
- The functional API allows user to create the same models but offers more flexibility at the cost of simplicity and readability. It can be used with multiple input and output layers and shared layers, which enables to build complex network structures. When using the functional API, we always need to pass the previous layer to the current layer. It also requires the use of an input layer. The functional API is also often used for transfer learning.
- Keras provides seven different datasets, which can be loaded using Keras directly. These include image datasets, a house price, and a movie review dataset. Keras is a high-level neural networks API running on top of Tensorflow.

- Demerits of Keras :
  1. Restricted to use only on smaller datasets.
  2. Slower performance since it is completely written in Python.
  3. The project repository is not as huge compared to TensorFlow.

1.13.3 Difference Between Keras and Tensorflow

Keras	Tensorflow
Keras is a high-level API which is running on top of TensorFlow.	TensorFlow is a framework that offers both high and low-level APIs.
Keras is used for small datasets as it is slower.	TensorFlow is used for high-performance models and massive datasets that require execution fast.
It can be used for low-performance models.	It is use for high-performance models.
Working with beginner-friendly projects with small datasets.	Rendering heavy projects with ease.
Readable and concise architecture	Comparatively complex architecture.
Community support is minimal but growing.	Being one of the first deep learning frameworks that were developed, it has an extensive community and a rich reserve of readable material.

1.13.4 PyTorch

- PyTorch is an open source machine learning library for Python and is completely based on Torch. It is primarily used for applications such as natural language processing. It was developed by Facebook's AI Research lab.
- PyTorch uses dynamic computation, which allows greater flexibility in building complex architectures. Pytorch uses core Python concepts like classes, structures and conditional loops.
- PyTorch has a very simple interface like Python. It provides an easy way to use API.
- It allows developers to train a neural network model in a distributed manner. It provides optimized performance in both research and production with the help of native support for peer to peer communication and asynchronous execution of collective operation from Python and C++.

- The PyTorch core is used to implement tensor data structure, CPU and GPU operators, basic parallel primitives and automatic differentiation calculations.
- PyTorch is supported by many major cloud platforms such as AWS. With the help of prebuilt images, large scale training on GPU's and ability to run models in a production scale environment etc.; it provides frictionless development and easy scaling.

1.13.5 Caffe

- Convolutional Architecture for Fast Feature Embedding (Caffe) is a deep learning framework made with expression, speed, and modularity in mind. It is developed by Berkeley AI Research (BAIR) and by community contributors.
- Caffe is primarily a C++ library and exposes a modular development interface. Caffe offers interfaces for daily use by way of the command line, Python and MATLAB.
- Caffe processes data in the form of Blobs which are N-dimensional arrays stored in a C - contiguous fashion. Data is stored both as data we pass along the model and as diff, which is a gradient computed by the network.
- Data layers handle how the data is processed in and out of the Caffe model. Pre-processing and transformation like random cropping, mirroring, scaling and mean subtraction can be done by configuring the data layer. Furthermore, pre-fetching and multiple-input configurations are also possible.

1.13.6 Shogun

- Shogun is an open source machine learning software library built in C++. Shogun was developed in 1999 by Soeren Sonnenburg and Gunnar Raetsch.
- It offers numerous algorithms and data structures for machine learning problems. The focus of Shogun is on kernel machines such as support vector machines for regression and classification problems. Shogun also offers a full implementation of Hidden Markov models.
- Shogun is a community learning platform in every sense, anyone can use the toolbox to learn about ML and apply it to solve problems.
- Shogun is community-based and non-commercial. It is currently released under the GPLv3. It is also moving towards BSD compatibility with optional GPL (General Public License) parts. This means companies can use the code without having to turn their code-base to GPL.

❑❑❑

Notes

Unit II

2 Deep Neural Networks

Syllabus

Introduction to Neural Networks : *The Biological Neuron, The Perceptron, Multilayer Feed-Forward Networks,* ***Training Neural Networks :*** *Backpropagation and Forward propagation* ***Activation Functions :*** *Linear, Sigmoid, Tannh, Hard Tanh, Softmax, Rectified Linear,* ***Loss Functions :*** *Loss Function Notation, Loss Functions for Regression, Loss Functions for Classification, Loss Functions for Reconstruction,* ***Hyperparameters :*** *Learning Rate, Regularization, Momentum, Sparsity, Deep Feedforward Networks - Example of Ex OR, Hidden Units, cost functions, error backpropagation, Gradient-Based Learning, Implementing Gradient Descent, vanishing and Exploding gradient descent, Sentiment Analysis, Deep Learning with Pytorch, Jupyter, colab.*

Contents

2.1 Introduction to Neural Networks

- Neural networks consists of many numbers of simple elements (neurons) connected between them in system. Whole system is able to solve of complex tasks and to learn for it like a natural brain.
- Neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.
- For user, NN is black box with input vector (source data) and output vector (result).
  - A Neural network is usually structured into an input layer of neurons, one or more hidden layers one output layer.
  - Neurons belonging to adjacent layers are usually fully connected and the various types and architectures are indentified both by the different topologies adopted for the connections as well as by the choice of the activation function.
  - The values of the functions associated with the connections are called "weights".
  - The whole game of using NNs is in fact that, in order for the network to yield appropriate outputs for given inputs, the weight must be set to suitable values. The way this is obtained allows a further distinction among modes of operations.
  - A neural network is a processing device, either an algorithm or actual hardware, whose design was motivated by the design and functioning of human brains and components thereof.
  - Most neural networks have some sort of "training" rule whereby the weights of connections are adjusted on the basis of presented patterns.
  - In other words, neural networks "learn" from example, just like children learn to recognize dogs from examples of dogs, and exhibit some structural capability for generalization.
  - Neural networks normally have great potential for paralleism, since the computations of the components are independent of each other.
  - Neural networks are a different paradigm for computing :
    1. Von Neumann machines are based on the processing/memory abstraction of human information processing.
    2. Neural networks are based on the paralllel architecture of animal brains.

- Neural networks are a form of multiprocessor computer system, with
  a. Simple processing elements
  b. A high degree of interconnections
  c. Simple scalar messages
  d. Adaptive interaction between elements
- The advantages of neural networks are due to its adaptive and generalization ability.
  a) Neural networks are adaptive methods that can learn without any prior assumption of the underlying data.
  b) Neural network, namely the feed forward multilayer perception and radial basis function network have been proven to be universal functional approximations.
  c) Neural networks are non-linear model with good generalization ability.
- **Useful properties and capabilities of neural network.**
  1. **Nonlinearity :** An artificial neuron can be linear or nonlinear. A neural network, made up of an interconnection of nonlinear neurons, is itself nonlinear.
  2. **Adaptivity :** Neural networks have a built-in capability to adapt their synaptic weights to changes in the surrounding environment.
  3. **Contextual information :** Knowledge is represented by the very structured and activation state of a neural network.
  4. **Evidential response :** In the context of pattern classification, a neural network can be designed to provide information not only about which particular pattern to select, but also about the confidence in the decision made.
  5. **Uniformity of analysis and design :** Neural networks enjoy universality as information processors.
  6. **VLSI implement-ability :** The massively parallel nature of a neural network makes it potentially fast for the computation of certain tasks.

2.1.1 Advantages of Neural Network

- The advantages of neural networks are due to its adaptive and generalization ability.
  a) Neural networks are adaptive methods that can learn without any prior assumption of the underlying data.
  b) Neural network, namely the feed forward multilayer perception and radial basis function network have been proven to be universal functional approximations.
  c) Neural networks are non-linear model with good generalization ability.

2.1.2 Application of Neural Network

- Neural network applications can be grouped in following categories :
  1. **Clustering :** A clustering algorithm explores the similarity between patterns and places similar patterns in a cluster. Best known applications include data compression and data mining.
  2. **Classification/Pattern recognition :** The task of pattern recognition is to assign an input pattern (like handwritten symbol) to one of many classes. This category includes algorithmic implementations such as associative memory.
  3. **Function approximation :** The tasks of function approximation is to find an estimate of the unknown function f() subject to noise. Various engineering and scientific disciplines require function approximation.
  4. **Prediction/Dynamical systems :** The task is to forecast some future values of a time-sequenced data. Prediciton has a significant impact on decision support systems. Prediction differs from function approximation by considering time factor.

2.1.3 Difference between Digital Computer and Neural Networks

Sr. No.	Digital Computers	Neural Networks
1.	Deductive reasoning : We apply known rules input data to produce output.	Inductuve reasoning : Given input and output data (training examples), we construct the rules.
2.	Computation is centralized, synchronous and serial.	Coputation is collectivem asynchronous and parallel.
3.	Memory is packetted, literally stored and location addressable.	Memory is distributed, internalized and content addressable.
4.	Not fault toerant. One transistor goes and it no longer works.	Fault tolerant, redundancy and sharing of responsibilities.
5.	Fast. Measured in millionths of a second.	Slow. Measured in thousands of a second.
6.	Exact.	Inexact.
7.	Static connectivity.	Dynamic connectivity.
8.	Applicable if well defined rules with precise input data.	Applicable if rules are unknown or complicated or if data is noisy or partial.

2.2 The Biological Neuron

- Artificial neural systems are inspired by biological neural systems. The elementary building block of biological neural systems is the neuron.
- The brain is a collection of about 10 billion interconnected neurons. Each neuron is a cell [right] that uses biochemical reactions to receive, process and transmit information. Fig. 2.2.1 shows biological neural systems.

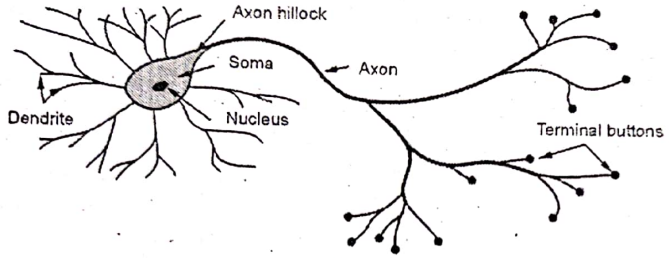


Fig. 2.2.1 Schematic of biological neuron

- The single cell neuron consists of the cell body or soma, the dendrites and the axon. The dendrites receive signals from the axons of other neurons. The small space between the axon of one neuron and the dendrite of another is the synapse. The afferent dendrites conduct impulses toward the soma. The efferent axon conducts impulses away from the soma.

Basic Components of Biological Neurons

1. The majority of neurons encode their activations or outputs as a series of brief electrical pulses (i.e. spikes or action potentials).
2. The neuron's ***cell body*** **(*soma*)** processes the incoming activations and converts them into output activations.
3. The neuron's ***nucleus*** contains the genetic material in the form of DNA. This exists in most types of cells, not just neurons.
4. ***Dendrites*** are fibres which emanate from the cell body and provide the receptive zones that receive activation from other neurons.
5. Axons are fibres acting as transmission lines that send activation to other neurons.
6. The junctions that allow signal transmission between the axons and dendrites are called ***synapses***. The process of transmission is by diffusion of chemicals called ***neurotransmitters*** across the synaptic cleft.

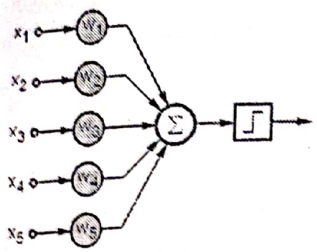
- Comparison between Biological NN and Artificial NN

Biological NN	Artificial NN
soma	unit
Axon, dendrite	dendrite
synapse	weight
potential	weighted sum
threshold	bias weight
signal	activation

2.3 The Perceptron

- A learning algorithm is an adaptive method by which a network of computing units self-organizes to implement the desired behaviour.
- Learning algorithms can be divided into supervised and unsupervised methods. Supervised learning denotes a method in which some input vectors are collected and presented to the network. The output computed by the network is observed and the deviation from the expected answer is measured.
- Unsupervised learning is used when, for a given input, the exact numerical output a network should produce is unknown. The perceptron learning algorithm is an example of supervised learning with reinforcement
- Rosenblatt's perceptron is built around a nonlinear neuron, namely, the McCulloch-Pitts model of a neuron.
- Rosenblatt perceptron is a binary single neuron model. The inputs integration is implemented through the addition of the weighted inputs that have fixed weights obtained during the training stage. If the result of this addition is larger than a given threshold ? the neuron fires. When the neuron fires its output is set to 1, otherwise it's set to 0.

- The equation can be re-written as follows including what it's known as the bias term :



$$h^{(x)} = \begin{cases} 1 & \text{if } w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d \geq \theta \\ 0 & \text{if } w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d < \theta \end{cases}$$

The equation can be re-written as follows including what its known as the bias term : $x_0 = 1$, $w_0 = \theta$.

$$h^{(x)} = \begin{cases} 1 & \text{if } w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d \geq 0 \\ 0 & \text{if } w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d < 0 \end{cases}$$

$$h^{(x)} = \begin{cases} 1 & \text{if } w^t \cdot x \geq 0 \\ 0 & \text{if } w^t \cdot x \geq 0 \end{cases}$$

2.3.1 ADALINE Network

- ADALINE (Adaptive Linear Neuron) is an early single-layer artificial neural network.
- An important generalized of the perceptrons training algorithm was presented by Widrow and Hoff as the least mean square learning procedure also known as the delta rule.
- The learning rule was applied to the "adaptive linear element" also named Adaline.
- The perceptron learning rule uses the output of the thersold function for learning. The delta rule uses the net output without further mapping into output values –1 or + 1.
- Fig. 2.3.1 shows Adaline.

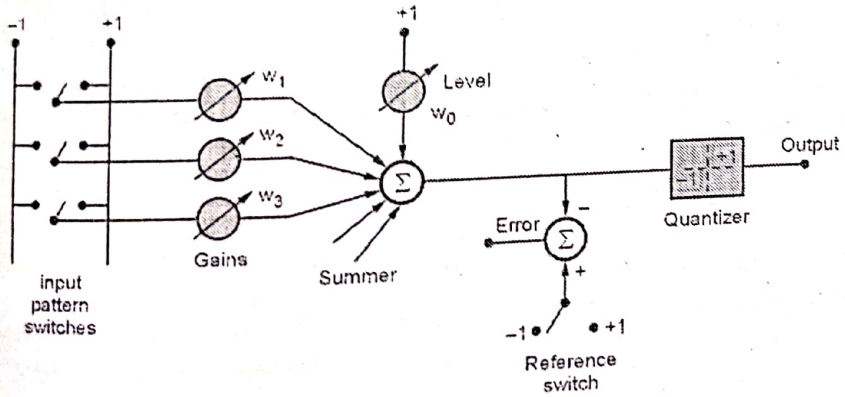


Fig. 2.3.1 Adaline

- If the input conductances are denoted by W_i where m i = 0, 1, 2, ..., n and input and output signals by x_{in} and y respectively, then the output of the central block is defined to be :

$$y = \sum_{i=1}^{x_{in}} w_i x_i + \theta$$

Where, $\theta = W_o$

- In a simple physical implementation, this device consists of a set of controllable resistors connected to a circuit which can sum up currents caused by the input voltage signals. Usually the central block, the summer is also followed by a quantizer which outputs + 1 of – 1, depending on the polarity of the sum.
- The problem is to determine the coefficients , where i = 0, 1 ..., n , in such way that the input output response is correct for a large number of arbitrarily chosen signal sets.
- If an exact mapping is not possible the average error must be minimized, for instance, in the sense of least squares.
- An adaptive operation means that there exists a mechanism by which the can be adjusted, usually iteratively to attain the correct values.
- For the Adaline, Widrow introduced the delta rule to adjust the weights.
- For the p^{th} input-output pattern, the earor measure of a single-output Adaline can be expressed as,

$$E_P = (t_p - o_p)^2$$

Where t_p = Target output

o_p = Actual output of the Adaline E_P

- The derivation of E_P with respect to each weight W_i is

$$\frac{\partial E_P}{\partial W_i} = -2 (t_p - o_p) x_i$$

- To decrease by gradient descent, the update formula for W_i on the p^{th} input-output pattern is

$$\Delta_P W_i = \eta (t_p - O_p) x_i$$

- The delta rule tries to minimize squared errors, it is also referred to as the least mean square learning procedure or Widrow - Hoff learning rule.

2.3.2 Sigmoid Neuron

- A sigmoid function produces a curve with an "S" shape. The example sigmoid function shown on the left is a special case of the logistic function, which models the growth of some set.

$$\text{sig}(t) = \frac{1}{1+e^{-t}}$$

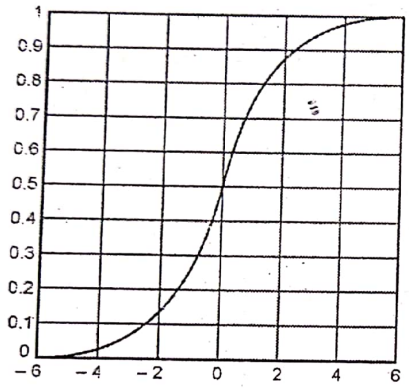


Fig. 2.3.2

- In general, a sigmoid function is real-valued and differentiable, having a non-negative or non-positive first derivative, one local minimum, and one local maximum.
- The logistic sigmoid function is related to the hyperbolic tangent as follows :

$$1 - 2\,\text{sig}(x) = 1 - 2\,\frac{1}{1+e^{-x}} = -\tan h\,\frac{x}{2}$$

- Sigmoid functions are often used in artificial neural networks to introduce nonlinearity in the model.
- A neural network element computes a linear combination of its input signals, and applies a sigmoid function to the result.
- A reason for its popularity in neural networks is because the sigmoid function satisfies a property between the derivative and itself such that it is computationally easy to perform.

$$\frac{d}{dt}\,\text{sig}(t) = \text{sig}(t)\,(1 - \text{sig}(t))$$

- Derivatives of the sigmoid function are usually employed in learning algorithms.

2.4 Architecture of Neural Networks

2.4.1 Single Layer Feed Forward Network

- The architecture of the neural network refers to the arrangement of the connection between neurons, processing element, number of layers, and the flow of signal in the neural network.
- There are mainly two category of neural network architecture :
  a. Feed-forward
  b. Feedback (recurrent) neural networks.

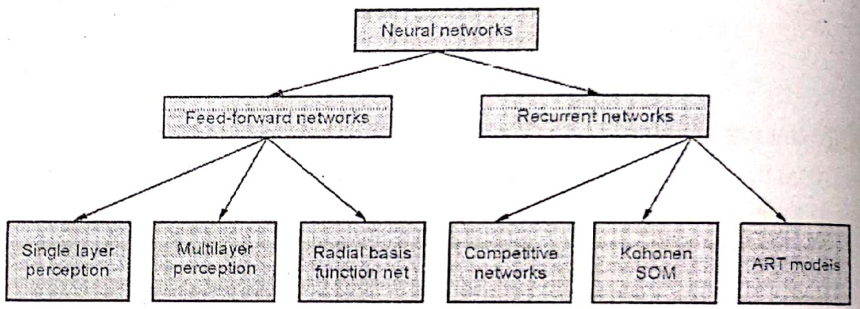


Fig. 2.4.1

1. Architecture and Learning Rule

- In late 1950s, Frank Rosenblatt introduced a network composed of the units that were enhanced version of McCulloch-Pitts Threshold Logic Unit (TLU) model.
- Rosenblatt's model of neuron, a perceptron, was the result of merger between two concepts from the 1940s, McCulloch-Pitts model of an artificial neuron and Hebbian learning rule of adjusting weights.
- In addition to the variable weight values, the perceptron model added an extra input that represents bias. Thus, the modified equation is now as follows :

$$\text{Sum} = \sum_{i=1}^{N} I_i\,W_i + b\,,$$

where b represents the bias value.

- Fig. 2.4.2 shows a typical perception setup for pattern recognition applications, in which visual patterns are represented as matrices of elements between 0 and 1.

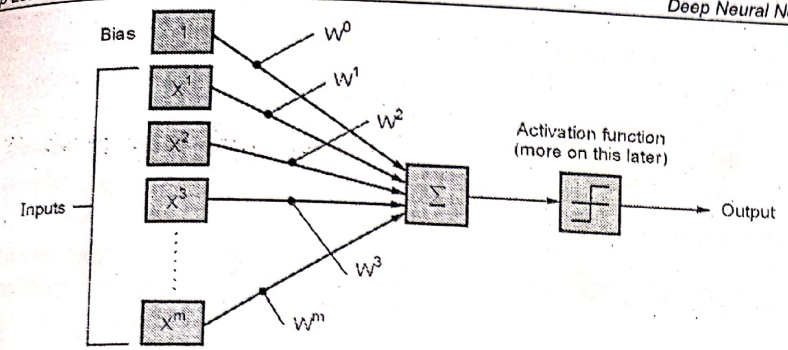


Fig. 2.4.2 Perception setup

1. First layer act as a set of feature detectors that are hardwired to the input signals to detect specific features.
2. Second layer i.e. output layer takes the outputs of the feature detectors in the first layer and classifies the given input pattern.

- Learning is initiated by making adjustments to the relevant connection strengths and a threshold value θ.
- Here we consider only two class problem. Here output layer usually has only a single node. For an n-class problem (n > 3), the output layer usually has n-nodes, each corresponding to a class and the output node with the largest value indicates which class the input vector belongs to.
- In the first stage, the linear combination of inputs is calculated. Each value of input array is associated with its weight value, which is normally between 0 and 1. Also, the summation function often takes an extra input value Theta with weight value of 1 to represent threshold or bias of a neuron.
  a. The term x_i is referred to as **active or excitatory** if its value is 1.
  b. If the value is 0 then it is **inactive.**
  c. If the value is – 1 then it is **inhibitory.**
- The output unit is a linear threshold element with a threshold value θ :

$$0 = f\left(\sum_{i=1}^{n} w_i x_i - \theta\right)$$

$$= f\left(\sum_{i=1}^{n} w_i x_i + w_0\right),\ w_0 = -\theta$$

$$= f\left(\sum_{i=1}^{n} w_i x_i\right),\ x_0 \equiv 1$$

where w_i is a modifiable weight associated with an incoming signal x_i.

- Fig. 2.4.3 shows the bias term w_0.

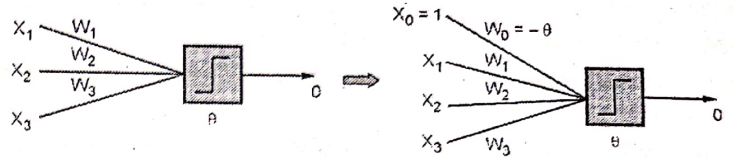


Fig. 2.4.3 Bias term W_0

- The function y = f(x) describes relationship, an input-output mapping from x to y.
- The equation (2.4.1), the f(.) is the **activation function** of the perceptron and it is typically either a **signum function** sgn(x) or **step function** step(x) :

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{otherwise} \end{cases}$$

$$\text{step}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases} \quad \ldots (2.4.1)$$

- The sum-of-product value is then passed into the second stage to perform the activation function which generates the output from the neuron. The activation function "squashes" the amplitude of the output in the range of [0, 1] or [– 1, 1] alternately. The behavior of the activation function will describe the characteristics of an artificial neuron model.
- The basic learning algorithm for a single layer perceptron repeats the following steps until the weights converge :
  1. Select an input vector x from the training data set.
  2. If the perceptron gives an incorrect response, modify all connection weights w_i according to

$$\Delta w_i = \eta\, t_i\, X_i$$

Where t_i is a target output and η is a learning state.

Perceptron Convergence Theorem

Theorem : If there is a set of weights that correctly classify the (linearly separable) training patterns, then the learning algorithm will find one such weight set, w^* in a finite number of iterations.

Assumptions :

1. At least one such set of weights, w^*, exists and

2. There are a finite number of training patterns.
3. The threshold function is uni-polar (output is 0 or 1).

2. Exclusive OR problem

- XOR problem is a pattern recognition problem in neural network.
- Neural networks can be used to classify boolean functions depending on their desired outputs.
- For a two input binary XOR problem, the desired output is given in the form of truth table.

	X	Y	Class
Desired I/O pair 1	0	0	0
Desired I/O pair 2	0	1	1
Desired I/O pair 3	1	0	1
Desired I/O pair 4	1	1	0

- The XOR problem is not **linearly separable**. We cannot use a single layer perceptron to construct a straight line to partition the two dimensional input space into two regions, each containing only data points of the same class.
- Let us consider following four equations :

$0 \times w_1 + 0 \times w_2 + w_0 \leq 0 \Leftrightarrow w_0 \leq 0,$

$0 \times w_1 + 1 \times w_2 + w_0 > 0 \Leftrightarrow w_0 > -w_2,$

$1 \times w_1 + 0 \times w_2 + w_0 > 0 \Leftrightarrow w_0 > -w_1,$

$1 \times w_1 + 1 \times w_2 + w_0 \leq 0 \Leftrightarrow w_0 \leq -w_1 - w_2$

2.4.2 Multi-Layer Feed Forward Network

- A multilayer feed-forward neural network is a network consisting of multiple layers of units, all of which are adaptive. The network is not allowed to have cycles from later layers back to earlier layers, hence the name "feed-forward". Let us consider a network with a single complete hidden layer. i.e., the network consists of some input nodes, some output nodes, and a set of hidden nodes. Every hidden node takes inputs from each of the input nodes and feeds into each of the output nodes.

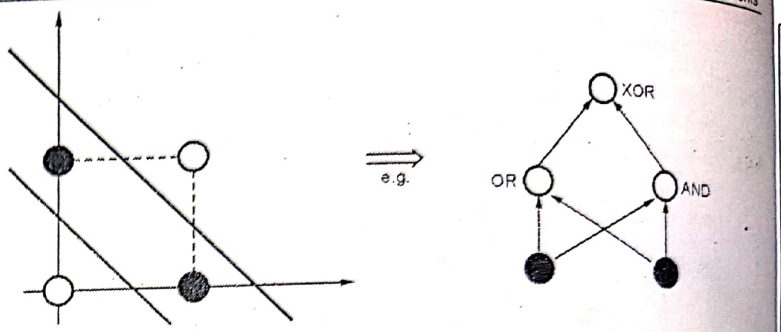


Fig. 2.4.4

- In multi-layer feed forward neural networks, the sigmoid activation function, defined by $g(x) = \frac{1}{1+e^{-x}}$ is normally used.
- A Multi-Layer Perceptron (MLP) has the same structure of a single layer perceptron with one or more hidden layers. An MLP is a network of simple neurons called perceptrons.
- A typical multilayer perceptron network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes and an output layer of nodes.
- It is not possible to find weights which enable single layer perceptrons to deal with non-linearly separable problems like XOR :
- Multi-layer perceptrons are able to cope with non-linearly separable problems.
- Each neuron in one layer has direct connections to all the neurons of the subsequent layer. MLP can implement nonlinear discriminants (for classification) and nonlinear regression functions (for regression).
- Historically, the problem was that there were no known learning algorithms for training MLPs. Fortunately; it is now known to be quite straightforward. The procedure for finding a gradient vector in the network structure is generally referred to as **backpropagation**. Because the gradient vector is calculated in the direction opposite to the flow of the output of each node.
- Procedure of backpropagation :
  1. The output values are compared with the target to compute the value of some predefined error function.
  2. The error is then fedback through the network.

3. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function.

- Continue this process until the connection weights in the network have been adjusted so that the network output has converged, to an acceptable level, with the desired output.
- If we use the gradient vector in a simple steepest descent method, the resulting learning paradigm is often referred to as the **backpropagation** learning rule. Backpropagation works by approximating the non-linear relationship between the input and the output by adjusting the weight values internally.
- Generally, the backpropagation network has two stages, training and testing. During the training phase, the network is "shown" sample inputs and the correct classifications. For example, the input might be an encoded picture of a face, and the output could be represented by a code that corresponds to the name of the person.
- Fig. 2.4.5 shows three most commonly used activation functions in backpropagation MLPs.

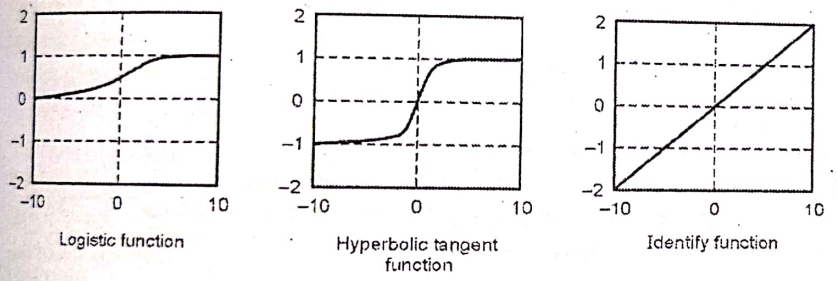


Fig. 2.4.5 Activation function

Logistic function :

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

Hyperbolic tangent function :

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

Identity function :

$$f(x) = x$$

- Both the hyperbolic tangent function and logistic function approximate the signum and step function respectively. Sometimes these two function are referred to as **squashing functions** since the inputs to these functions are squashed to the range [0, 1] or [– 1, 1].
- These functions are also called **sigmoidal functions** because their S-shaped curves exhibits smoothness and asymptotic properties.
- A learning process is organized through a learning algorithm, which is a process of updating the weights in such a way that a machine learning tool implements a given input/output mapping with no errors or with some minimal acceptable error.
- Any learning algorithm is based on a certain learning rule, which determines how the weights shall be updated if the error occurs.

Backpropagation Learning Rule

- The **net input** of a node is defined as the weighted sum of the incoming signals plus a bias term. Fig. 2.4.6 shows the backpropagation MLP for node j. The net input and output of node j is as follows :

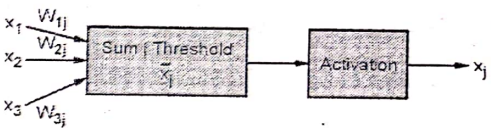


Fig. 2.4.6 Backpropagation MLP for node j

$$\bar{X}_j = \Sigma_i + W_{ij} + W_j$$

$$x_j = f(\bar{X}_j) = \frac{1}{1 + \exp(-\bar{X}_j)}$$

Where

x_i is the output of node i located in any one of the pervious layers,

W_{ij} is the weight associated with the link connecting nodes i and j,

W_j is the bias of node j.

- Internal parameters associated with each node j is the weight W_{ij}. So changing the weights of the node will change the behaviour of the whole backpropagation MLP.
- Fig. 2.4.7 shows two layer backpropagation MLP.
- The above backpropagation MLP will refer to as a 3-4-3 network, corresponding to the number of nodes in each layer.
- The backward error propagation also known as the backpropagation (BP) or the

Generalized Delta Rule (GDR). A squared error measure for the p^{th} input-output pair is defined as

$$E_p = \sum_k (d_k - x_k)^2$$

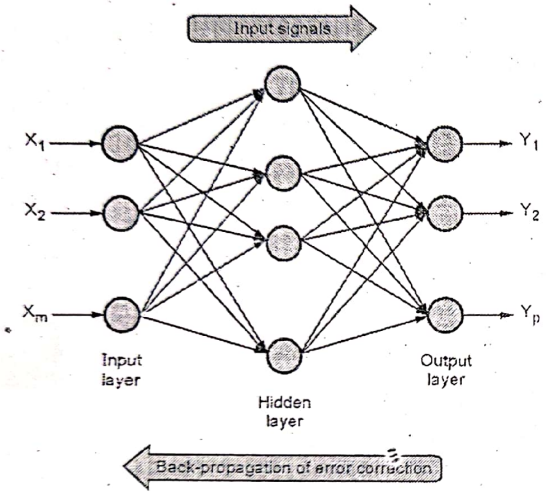


Fig. 2.4.7 Two layer backpropagation MLP

Where d_k is the desired output for node k and x_k is the actual output for node k when the input part of the p^{th} data pair is presented.

- To find the gradient vector, an error term for node i is defined as :

$$\bar{\epsilon}_i = \frac{\partial + E_p}{\partial \bar{X}_i}$$

- The partial derivative can be rewritten as product of two terms using chain rule for partial differentiation :

$$\frac{\partial E(t)}{\partial w_{ij}(t)} = \frac{\partial E(t)}{\partial a_i(t)} \cdot \frac{\partial a_i(t)}{w_{ij}(t)}$$

- Features of the delta rule are as follows :
  1. Simplicity
  2. Distributed learning : Learning is not reliant on central control of the network.
  3. Online learning : Weights are updated after presentation of each pattern.

Rules for Feedforward Multilayer Perceptron

- The training algorithm is called Error Back Propagation (EBP) training algorithm. If a submitted pattern provides an output far from desired value, the weights and thresholds are adjusted so that the current mean square classification error is reduced.
- The training is repeated for all patterns until the training set provide an acceptable overall error. Usually the mapping error is computed over the full training set.
- Error back propagation algorithm is working in two stages :
  1. The trained network operates feed-forward to obtain output of the network
  2. The weight adjustment propagate backward from output layer through hidden layer toward input layer.

2.4.3 Recurrent Neural Network

- A recurrent neural network is a type of neural network that contains loops, allowing information to be stored within the network.
- A RNN is particularly useful when a sequence of data is being processed to make a classification decision or regression estimate but it can also be used on non-sequential data. Recurrent neural networks are typically used to solve tasks related to time series data.
- Applications of recurrent neural networks include natural language processing, speech recognition, machine translation, character-level language modeling, image classification, image captioning, stock prediction and financial engineering.

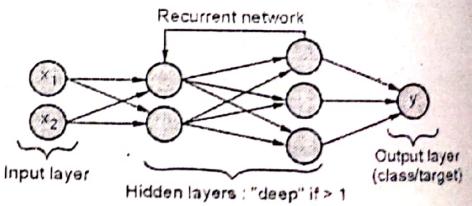


Fig. 2.4.8 Architecture of RNN

- Fig. 2.4.8 shows architecture of recurrent neural network.
- Recurrent Neural Networks can be thought of as a series of networks linked together. They often have a chain-like architecture, making them applicable for tasks such as speech recognition, language translation, etc.
- An RNN can be designed to operate across sequences of vectors in the input, output, or both. For example, a sequenced input may take a sentence as an input and output a positive or negative sentiment value. Alternatively, a sequenced output may take an image as an input and produce a sentence as an output.

2.5 Training Neural Networks

- Well-trained artificial neural network has weights that amplify the signal and creates the noise. A bigger weight signifies a tighter correlation between a signal and the network's outcome. Inputs paired with large weights will affect the network's interpretation of the data more than inputs paired with smaller weights.
- The process of learning for any learning algorithm using weights is the process of readjusting the weights and biases. In most datasets, certain features are strongly correlated with certain labels. Neural networks learn these relationships blindly by making a guess based on the inputs and weights.
- The loss functions in optimization algorithms, such as stochastic gradient descent (SGD), reward the network for good guesses and penalize it for bad ones.

2.5.1 Backpropagation

- Backpropagation is a training method used for a multi-layer neural network. It is also called the generalized delta rule. It is a gradient descent method which minimizes the total squared error of the output computed by the net.
- The backpropagation algorithm looks for the minimum value of the error function in weight space using a technique called the delta rule or gradient descent. The weights that minimize the error function is then considered to be a solution to the learning problem.
- Backpropagation is a systematic method for training multiple layer ANN. It is a generalization of Widrow-Hoff error correction rule. 80 % of ANN applications uses backpropagation.
- Fig. 2.5.1 shows backpropagation network.

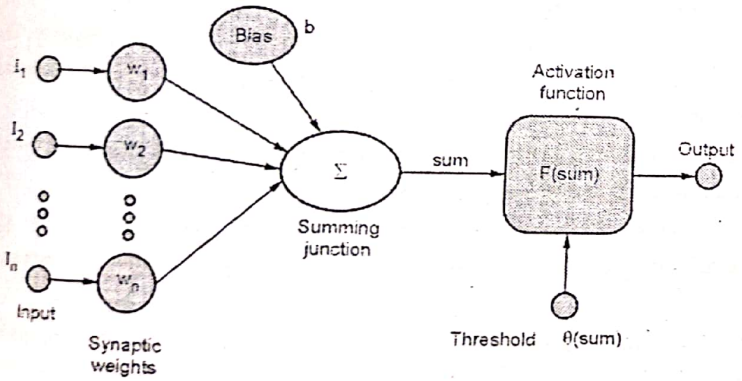


Fig. 2.5.1 Backpropagation network

- Consider a simple neuron :
  a. Neuron has a summing junction and activation function.
  b. Any non linear function which differentiable everywhere and increases everywhere with sum can be used as activation function.
  c. Examples : Logistic function, Arc tangent function, Hyperbolic tangent activation function.
- These activation function makes the multilayer network to have greater representational power than single layer network only when non-linearity is introduced.
- **Need of hidden layers :**
  1. A network with only two layers (input and output) can only represent the input with whatever representation already exists in the input data.
  2. If the data is discontinuous or non-linearly separable, the innate representation is inconsistent, and the mapping cannot be learned using two layers (Input and Output).
  3. Therefore, hidden layer(s) are used between input and output layers
- **Weights** connects unit (neuron) in one layer only to those in the next higher layer. The output of the unit is scaled by the value of the connecting weight, and it is fed forward to provide a portion of the activation for the units in the next higher layer.
- Backpropagation can be applied to an artificial neural network with any number of hidden layers. The training objective is to adjust the weights so that the application of a set of inputs produces the desired outputs.
- **Training procedure :** The network is usually trained with a large number of input - output pairs.
  1. Generate weights randomly to small random values (both positive and negative) to ensure that the network is not saturated by large values of weights.
  2. Choose a training pair from the training set.
  3. Apply the input vector to network input.
  4. Calculate the network output.
  5. Calculate the error, the difference between the network output and the desired output.
  6. Adjust the weights of the network in a way that minimizes this error.
  7. Repeat steps 2 - 6 for each pair of input-output in the training set until the error for the entire system is acceptably low.

Forward pass and backward pass :

- Backpropagation neural network training involves two passes.
  1. In the forward pass, the input signals moves forward from the network input to the output.
  2. In the backward pass, the calculated error signals propagate backward through the network, where they are used to adjust the weights.
  3. In the forward pass, the calculation of the output is carried out, layer by layer, in the forward direction. The output of one layer is the input to the next layer.
- In the reverse pass,
  a. The weights of the output neuron layer are adjusted first since the target value of each output neuron is available to guide the adjustment of the associated weights, using the delta rule.
  b. Next, we adjust the weights of the middle layers. As the middle layer neurons have no target values, it makes the problem complex.
- **Selection of number of hidden units :** The number of hidden units depends on the number of input units.
  1. Never choose h to be more than twice the number of input units.
  2. You can load p patterns of I elements into log2 p hidden units.
  3. Ensure that we must have at least 1/e times as many training examples.
  4. Feature extraction requires fewer hidden units than inputs.
  5. Learning many examples of disjointed inputs requires more hidden units than inputs.
  6. The number of hidden units required for a classification task increases with the number of classes in the task. Large networks require longer training times.

Factors Influencing Backpropagation training

- The training time can be reduced by using :
  1. **Bias :** Networks with biases can represent relationships between inputs and outputs more easily than networks without biases. Adding a bias to each neuron is usually desirable to offset the origin of the activation function. The weight of the bias is trainable similar to weight except that the input is always + 1.
  2. **Momentum :** The use of momentum enhances the stability of the training process. Momentum is used to keep the training process going in the same general direction analogous to the way that momentum of a moving object

behaves. In backpropagation with momentum, the weight change is a combination of the current gradient and the previous gradient.

2.5.2 Advantages and Disadvantages of Backpropagation

Advantages of backpropagation :

1. It is simple, fast and easy to program.
2. Only numbers of the input are tuned and not any other parameter.
3. No need to have prior knowledge about the network.
4. It is flexible.
5. A standard approach and works efficiently.
6. It does not require the user to learn special functions.

Disadvantages of backpropagation :

1. Backpropagation possibly be sensitive to noisy data and irregularity.
2. The performance of this is highly reliant on the input data.
3. Needs excessive time for training.
4. The need for a matrix-based method for backpropagation instead of mini - batch.

2.6 Activation Functions

- Activation functions also known as transfer function is used to map input nodes to output nodes in certain fashion.
- The activation function is the most important factor in a neural network which decided whether or not a neuron will be activated or not and transferred to the next layer.
- Activation functions help in normalizing the output between 0 to 1 or – 1 to 1. It helps in the process of backpropagation due to their differentiable property. During backpropagation, loss function gets updated, and activation function helps the gradient descent curves to achieve their local minima.
- Activation function basically decides in any neural network that given input or receiving information is relevant or it is irrelevant.
- These activation function makes the multilayer network to have greater representational power than single layer network only when non-linearity is introduced.
- The input to the activation function is sum which is defined by the following equation.

$$\text{sum} = I_1W_1 + I_2W_2 + \ldots + I_nW_n = \sum_{j=1}^{n} I_j W_j + b$$

- **Activation Function : Logistic Function**

$$f(\text{sum}) = \frac{1}{(1 + e^{-s * \text{sum}})}$$

$$= (1 + s^{-s \times \text{sum}}) - 1$$

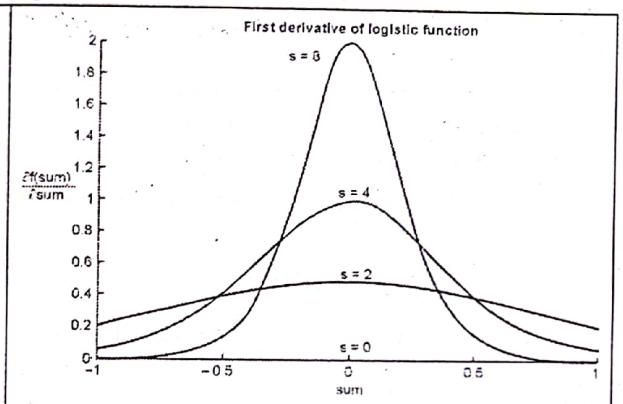


Fig. 2.6.1

- Logistic function monotonically increases from a lower limit (0 or – 1) to an upper limit (+ 1) as sum increases. In which values vary between 0 and 1, with a value of 0.5 when I is zero.
- **Activation Function : Arc Tangent**

$$f(\text{sum}) = \frac{2}{\pi} \tan^{-1} (s \times \text{sum})$$

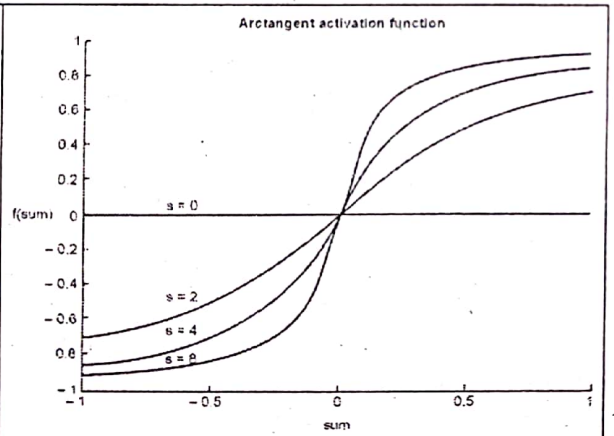


Fig. 2.6.2

- **Activation Function : Hyperbolic Tangent**

$$f(\text{sum}) = \tanh (s * I)$$

$$= \frac{e^{s \times \text{sum}} - e^{-s \times \text{sum}}}{e^{s \times \text{sum}} + e^{-s \times \text{sum}}}$$

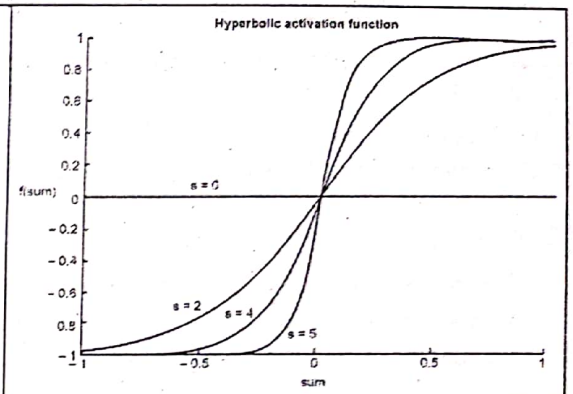


Fig. 2.6.3

2.6.1 Identity or Linear Activation Function

- A linear activation is a mathematical equation used for obtaining output vectors with specific properties.
- It is a simple straight line activation function where our function is directly proportional to the weighted sum of neurons or input.
- Linear activation functions are better in giving a wide range of activations and a line of a positive slope may increase the firing rate as the input rate increases.
- Fig. 2.6.4 shows identity function.

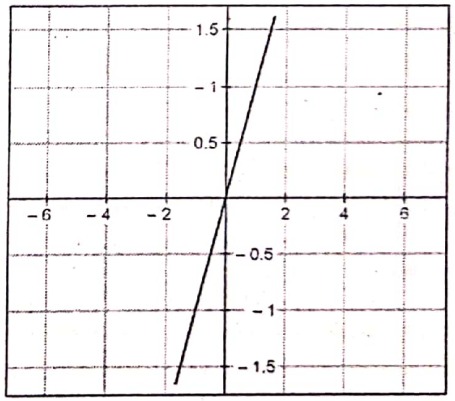


Fig. 2.6.4

2.7.1 Loss Function Notation

- Loss function notation are as follows :
  a) N = the number of samples collected.
  b) P = the number of input features gathered
  c) M = the number of output features that have been observed.
  d) (X, Y) to denote the input and output data collected. there will be N such pairs where the input is a collection of P values and the output Y is a collection of M values. We will denote the i^{th} pair in the dataset as X_i and Y_i.
  e) $\hat{Y}$ = output of the neural net.
  f) $h(X_i) = \hat{Y}_i$ = neural network transforming the input X_i to give the output.
  g) Thus $y_{i,j}$ refers to the j^{th} feature observed in the i^{th} sample collected.
  h) Loss function = L(W, b)

2.7.2 Loss Functions for Regression

- Loss functions for regression : Regression involves predicting a specific value that is continuous in nature. Estimating the price of a house or predicting stock prices are examples of regression because one works towards building a model that would predict a real-valued quantity.

Mean Square Error :

- Mean Sqaure Error (MSE) is the most commonly used regression loss function. MSE is the sum of squared distances between our target variable and predicted values.
- Mean Squared Error is the average of the squared differences between the actual and the predicted values. For a data point Y_i and its predicted value $\hat{Y}_i$, where n is the total number of data points in the dataset, the mean squared error is defined as :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **Advantages :** For small errors, MSE helps converge to the minima efficiently, as the gradient reduces gradually.

- Drawback :
  a) Squaring the values does increases the rate of training, but at the same time, an extremely large loss may lead to a drastic jump during backpropagation, which is not desirable.
  b) MSE is also sensitive to outliers.

2.7.3 Loss Functions for Classification

- Loss functions for classification : Classification problems involve predicting a discrete class output. It involves dividing the dataset into different and unique classes based on different parameters so that a new and unseen record can be put into one of the classes.

1. Hinge Loss

- Hinge Loss is a specific loss function used by Support Vector Machines (SVM). This loss function will help SVM to make a decision boundary with a certain margin distance.
- The equation for hinge loss when data points must be categorized as – 1 or 1 is as follows :

$$L(W,b) = \frac{1}{N}\sum_{i=1}^{N}\max(0, 1 - y_{ij} \times \hat{y}_{ij})$$

- Hinge loss is mostly used of binary classification.

Square hinge loss :

- There are many extensions of hinge loss are present to use with SVM models. One of the popular extensions is called Squared Hinge Loss. It simply calculates the square of the hinge loss value.
- Squared hinge loss has the effect of the smoothing the surface of the error function and making it numerically easier to work with.
- When the hinge loss requires better performance on a given binary classification problem it is mostly observed that a squared hinge loss may be appropriate to use. As using the hinge loss function, the target variable must be modified to have values in the set {– 1, 1}.
- It is simple to implement using python only we have to change the loss function name to "squared_hinge" in compile () function when building the model.
- A typical application can be classifying email into 'spam' and 'not spam' and we are only interested in the classification accuracy. Let us see how squared Hinge can be

used with Keras. It just involves specifying it as the used loss function during the model compilation step :

```
#Compile the model
model.compile(loss=squared_hinge,optimizer=tensorflow.keras.optimizers.Adam(lr=0.03), metrics=['accuracy'])
```

2.8 Hyper Parameters

- Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning.
- While designing a machine learning model, one always has multiple choices for the architectural design for the model. This creates a confusion on which design to choose for the model based on its optimality. And due to this, there are always trials for defining a perfect machine learning model.
- The parameters that are used to define these machine learning models are known as the hyperparameters and the rigorous search for these parameters to build an optimized model is known as hyperparameter tuning.
- Hyperparameters are not model parameters, which can be directly trained from data. Model parameters usually specify the way to transform the input into the required output, whereas hyperparameters define the actual structure of the model that gives the required data.

2.8.1 Layer Size

- Layer size is defined by the number of neurons in a given layer. Input and output layers are relatively easy to figure out because they correspond directly to how our modeling problem handles input and output.
- For the input layer, this will match up to the number of features in the input vector. For the output layer, this will either be a single output neuron or a number of neurons matching the number of classes we are trying to predict.
- It is obvious that a neural network with 3 layers will give better performance than that of 2 layers. Increasing more than 3 doesn't help that much in neural networks. In the case of CNN, an increasing number of layers makes the model better.

2.8.2 Magnitude : Learning Rate

- The amount that the weights are updated during training is referred to as the step size or the learning rate. Specifically, the learning rate is a configurable hyper-parameter used in the training of neural networks that has a small positive value, often in the range between 0.0 and 1.0.
- For example, if learning rate is 0.1, then the weights in the network are updated 0.1^* (estimated weight error) or 10 % of the estimated weight error each time the weights are updated. The learning rate hyper-parameter controls the rate or speed at which the model learns.
- Learning rates are tricky because they end up being specific to the dataset and even to other hyper-parameters. This creates a lot of overhead for finding the right setting for hyper-parameters.
- Large learning rates () make the model learn faster but at the same time it may cause us to miss the minimum loss function and only reach the surrounding of it. In cases where the learning rate is too large, the optimizer overshoots the minimum and the loss updates will lead to divergent behaviours.
- On the other hand, choosing lower learning rate values gives a better chance of finding the local minima with the trade-off of needing larger number of epochs and more time.
- **Momentum** can accelerate learning on those problems where the high-dimensional weight space that is being navigated by the optimization process has structures that mislead the gradient descent algorithm, such as flat regions or steep curvature.

2.9 Regularization

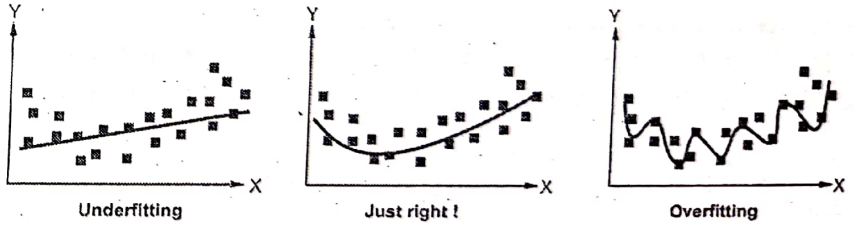


Fig. 2.9.1

- Just have a look at the above figure, and we can immediately predict that once we try to cover every minutest feature of the input data, there can be irregularities in

the extracted features, which can introduce noise in the output. This is referred to as "Overfitting".

- This may also happen with the lesser number of features extracted as some of the important details might be missed out. This will leave an effect on the accuracy of the outputs produced. This is referred to as "Underfitting".
- This also shows that the complexity for processing the input elements increases with overfitting. Also, neural networks being a complex interconnection of nodes, the issue of overfitting may arise frequently.
- To eliminate this, regularization is used, in which we have to make the slightest modification in the design of the neural network, and we can get better outcomes.

2.9.1 Regularization in Machine Learning

- One of the most important factors that affect the machine learning model is overfitting.
- The machine learning model may perform poorly if it tries to capture even the noise present in the dataset applied for training the system, which ultimately results in overfitting. In this context, noise doesn't mean the ambiguous or false data, but those inputs which do not acquire the required features to execute the machine learning model.
- Analyzing these data inputs may surely make the model flexible, but the risk of overfitting will also increase accordingly.
- One of the ways to avoid this is to cross validate the training dataset and decide accordingly the parameters to include that can increase the efficiency and performance of the model.
- Let this be the simple relation for linear regression :

$$Y \approx b_0 + b_1X_1 + b_2X_2 + \ldots b_pX_p$$

Where Y = Learned relation

β = Co-efficient estimators for different variables and/or predictors (X)

- Now, we shall introduce a loss function, that implements the fitting procedure, which is referred to as "Residual Sum of Squares" or RSS.
- The co-efficient in the function is chosen in such a way that it can minimize the loss function easily.

Hence,

$$RSS = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_jX_{ij}\right)^2$$

- Above equation will help in adjusting the co-efficient function depending on the training dataset.
- In case noise is present in the training dataset, then the adjusted co-efficient won't be generalized when the future datasets will be introduced. Hence, at this point, regularization comes into picture and makes this adjusted co-efficient shrink towards zero.
- One of the methods to implement this is the ridge regression, also known as L2 regularization. Lets have a quick overview on this.

2.9.2 Ridge Regression (L2 Regularization)

- Ridge regression, also known as L2 regularization, is a technique of regularization to avoid the overfitting in training data set, which introduces a small bias in the training model, through which one can get long term predictions for that input.
- In this method, a penalty term is added to the cost function. This amount of bias altered to the cost function in the model is also known as ridge regression penalty.
- Hence, the equation for the cost function, after introducing the ridge regression penalty is as follows :

$$\sum_{i=1}^{m}(y_i - y'_i)^2 = \sum_{i=1}^{m}\left(Y_i - \sum_{j=1}^{n}\beta_j \times X_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j^2$$

Here, λ is multiplied by the square of the weight set for the individual feature of the input data. This term is ridge regression penalty.

- It regularizes the co-efficient set for the model and hence the ridge regression term deduces the values of the coefficient, which ultimately helps in deducing the complexity of the machine learning model.
- From the above equation, we can observe that if the value of tends to zero, the last term on the right - hand side will tend to zero, thus making the above equation a representation of a simple linear regression model.
- Hence, lower the value of , the model will tend to linear regression.
- This model is important to execute the neural networks for machine learning, as there would be risks of failure for generalized linear regression models, if there are dependencies found between its variables. Hence, ridge regression is used here.

2.9.3 Lasso Regression (L1 Regularization)

- One more technique to reduce the overfitting and thus the complexity of the model is the lasso regression.
- Lasso regression stands for Least Absolute and Selection Operator and is also sometimes known as L1 regularization.
- The equation for the lasso regression is almost same as that of the ridge regression, except for a change that the value of the penalty term is taken as the absolute weights.
- The advantage of taking the absolute values is that its slope can shrink to 0, as compared to the ridge regression, where the slope will shrink it near to 0.
- The following equation gives the cost function defined in the Lasso regression :

$$\sum_{i=1}^{m}(y_i - y'_i)^2 = \sum_{i=1}^{m}\left(Y_i - \sum_{j=0}^{n}\beta_j \times X_{ij}\right)^2 + \lambda \sum_{j=0}^{n}|\beta_j|^2$$

- Due to the acceptance of absolute values for the cost function, some of the features of the input dataset can be ignored completely while evaluating the machine learning model and hence the feature selection and overfitting can be reduced to much extent.
- On the other hand, ridge regression does not ignore any feature in the model and includes it all for model evaluation. The complexity of the model can be reduced using the shrinking of co-efficient in the ridge regression model.

2.9.4 Dropout

- Dropout was introduced by "Hinton et al"and this method is now very popular. It consists of setting to zero the output of each hidden neuron in chosen layer with some probability and is proven to be very effective in reducing overfitting.
- Fig. 2.9.2 shows dropout regulations.
- To achieve dropout regularization, some neurons in the artificial neural network are randomly disabled. That prevents them from being too dependent on one another as they learn the correlations. Thus, the neurons work more independently, and the artificial neural network learns multiple independent correlations in the data based on different configurations of the neurons.
- It is used to improve the training of neural networks by omitting a hidden unit. It also speeds training.

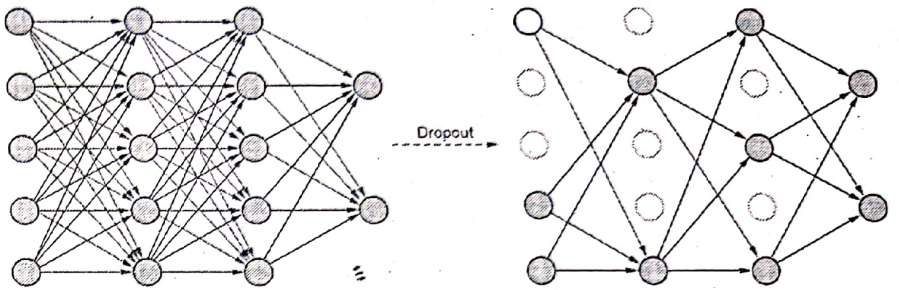


Fig. 2.9.2 Dropout regulation

- Dropout is driven by randomly dropping a neuron so that it will not contribute to the forward pass and back-propagation.
- Dropout is an inexpensive but powerful method of regularizing a broad family of models.

2.9.5 DropConnect

- DropConnect, known as the generalized version of Dropout, is the method used for regularizing deep neural networks. Fig. 2.9.3 shows dropconnect.

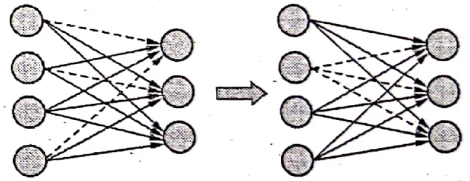


Fig. 2.9.3 Dropconnect

- DropConnect has been proposed to add more noise to the network. The primary difference is that instead of randomly dropping the output of the neurons, we randomly drop the connection between neurons.
- In other words, the fully connected layer with DropConnect becomes a sparsely connected layer in which the connections are chosen at random during the training stage.

2.9.6 Difference between L1 and L2 Regularization

L1 regulation	L2 regulation
Calculate the sum of the absolute values of the weights, called L1.	Calculate the sum of the squared values of the weights, called L2.
L1 has a sparse solution	L2 has a non sparse solution
It has built in feature selection	There is no feature selection
L1 is robust to outliers	L2 is not robust to outliers
L1 generates model that are simple and interpretable but cannot learn complex patterns	L2 regularization is able to learn complex data patterns
L1 has multiple solutions	L2 has one solution
It is equivalent to MAP Bayesian estimation with Laplace prior	It is equivalent to MAP Bayesian estimation with Gaussian prior.
A regression model that uses L1 regularization technique is called Lasso Regression	model which uses L2 is called Ridge Regression

2.10 Deep Feedforward Networks

- Deep feedforward networks is also called feedforward neural networks or multilayer perceptrons.
- Feedforward neural networks are called networks because they are typically represented by composing together many different functions. The model is associated with a directed acyclic graph describing how the functions are composed together.
- For example : consider 3 functions $f^{(1)}$, $f^{(2)}$ and $f^{(3)}$ connected in a chain, to form $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$.
- These chain structures are the most commonly used structures of neural networks. In this case, $f^{(1)}$ is called the **first layer** of the network, $f^{(2)}$ is called the **second layer** and so on.
- The overall length of the chain gives the depth of the model. It is from this terminology that the name "deep learning" arises.

- The final layer of a feedforward network is called the **output layer**. During neural network training, we drive $f(x)$ to match $f^*(x)$.
- The training data provides us with noisy, approximate examples of $f^*(x)$ evaluated at different training points.
- The training examples specify directly what the output layer must do at each point x; it must produce a value that is close to y. The behavior of the other layers is not directly specified by the training data.
- The learning algorithm must decide how to use those layers to produce the desired output, but the training data does not say what each individual layer should do.
- Instead, the learning algorithm must decide how to use these layers to best implement an approximation of f^*.
- Because the training data does not show the desired output for each of these layers, these layers are called **hidden layers**.
- Feedforward networks have introduced the concept of a hidden layer and this requires us to choose the activation functions that will be used to compute the hidden layer values.
- Finally, these networks are called neural because they are loosely inspired by neuroscience. Each hidden layer of the network is typically vector-valued. The dimensionality of these hidden layers determines the **width** of the model

2.10.1 Learning Concept of XOR

- XOR problem is a pattern recognition problem in neural network.
- The XOR function is an operation on two binary values, x_1 and x_2. When exactly one of these binary values is equal to 1, the XOR function returns 1. Otherwise, it returns 0.
- The XOR function provides the target function $y = f^*(x)$ that we want to learn. Our model provides a function $y = f(x; \theta)$ and our learning algorithm will adapt the parameters θ to make f as similar as possible to f^*.
- Neural networks can be used to classify Boolean functions depending on their desired outputs.
- The XOR problem is not **linearly separable**. We cannot use a single layer perceptron to construct a straight line to partition the two dimensional input space into two regions, each containing only data points of the same class.

2.11 Gradient-Based Learning

- Much of machine learning can be written as an optimization problem.
- Example loss functions : Logistic regression, linear regression, principle component analysis, neural network loss.
- A very efficient way to train logistic models is with **Stochastic Gradient Descent (SGD).**
- One challenge with training on power law data (i.e. most data) is that the terms in the gradient can have very different strengths
- The idea behind stochastic gradient descent is iterating a weight update based on the gradient of loss function :

$$\bar{w}(k+1) = \bar{w}(k) - \gamma \nabla L(\bar{w})$$

- Logistic regression is designed as a **binary classifer** (output say {0, 1}) but actually **outputs the probability** that the input instance is in the "1" class.
- A logistic classifier has the form :

$$p(X) = \frac{1}{1 + \exp(-X\beta)}$$

where $X = (X_1, \ldots, X_n)$ is a vector of features.

- Stochastic gradient has some serious limitations however, especially if the gradients vary widely in magnitude. Some coefficients change very fast, others very slowly.
- This happens for text, user activity and social media data (and other power-law data), because gradient magnitudes scale with feature frequency, i.e. over several orders of magnitude.
- It is not possible to set a single learning rate that trains the frequent and infrequent features at the same time.
- An example of stochastic gradient descent with perceptron loss is shown as follows :

from sklearn.linear_model import SGDClassifier

2.11.1 Finding the Optimal Hyper-parameters through Grid Search

- In statistics, hyperparameter is a parameter from a prior distribution; it captures the prior belief before data is observed.
- In any machine learning algorithm, these parameters need to be initialized before training a model.

- **Model hyperparameters** *are the properties that govern the entire training process.*
- Hyperparameters are important because they directly control the behaviour of the training algorithm and have a significant impact on the performance of the model is being trained.
- Choosing appropriate hyperparameters plays a crucial role in the success of our neural network architecture. Since it makes a huge impact on the learned model.
- For example, if the learning rate is too low, the model will miss the important patterns in the data. If it is high, it may have collisions.
- Choosing good hyperparameters gives two benefits :
  1. Efficiently search the space of possible hyperparameters.
  2. Easy to manage a large set of experiments for hyperparameter tuning.
- **The process of finding most optimal hyperparameters in machine learning is called hyperparameter optimisation.**
- Grid search is a very traditional technique for implementing hyperparameters. It brute force all combinations. Grid search requires to create two set of hyperparameters.

  1. Learning rate 2. Number of layers
- Grid search trains the algorithm for all combinations by using the two set of hyperparameters and measures the performance using "Cross Validation" technique.
- This validation technique gives assurance that our trained model got most of the patterns from the dataset.
- One of the best methods to do validation by using "K-Fold Cross Validation" which helps to provide ample data for training the model and ample data for validations.
- With this technique, we simply build a model for each possible combination of all of the hyperparameter values provided, evaluating each model and selecting the architecture which produces the best results.
- For example, say you have two continuous parameters α and β, where manually selected values for the parameters are the following :

$$\alpha \in \{0, 1, 2\}.$$
$$\beta \in \{.25, .50, .75\}$$

- Then the pairing of the selected hyperparametric values, H, can take on any of the following :

$$H \in \{(0, .25), (0, .50), (0, .75), (1, .25), (1, .50), (1, .75), (2, .25), (2, .50), (2, .75)\}$$

- Grid search will examine each pairing of and to determine the best performing combination. The resulting pairs, *H*, are simply each output that results from taking the Cartesian product of α and β.
- While straightforward, this "brute force" approach for hyperparameter optimization has some drawbacks. Higher-dimensional hyperparametric spaces are far more time consuming to test than the simple two-dimensional problem presented here.
- Also, because there will always be a fixed number of training samples for any given model, the model's predictive power will decrease as the number of dimensions increases. This is known as Hughes phenomenon.

2.11.2 Vanishing Gradient Problem

- When back-propagation is used, the earlier layers will receive very small updates compared to the later layers. This problem is referred to as the vanishing gradient problem.
- The vanishing gradient problem is essentially a situation in which a deep multilayer feed-forward network or a Recurrent Neural Network (RNN) does not have the ability to propagate useful gradient information from the output end of the model back to the layers near the input end of the model.
- Weight initialization is one technique that can be used to solve the vanishing gradient problem. It involves artificially creating an initial value for weights in a neural network to prevent the backpropagation algorithm from assigning weights that are unrealistically small.
- The most important solution to the vanishing gradient problem is a specific type of neural network called Long Short-Term Memory Networks (LSTMs).
- Indication of vanishing gradient problem :
  a) The parameters of the higher layers change to a great extent, while the parameters of lower layers barely change.
  b) The model weights could become 0 during training.
  c) The model learns at a particularly slow pace and the training could stagnate at a very early phase after only a few iterations.
- Some methods that are proposed to overcome the vanishing gradient problem:
  a) Residual neural networks (ResNets)
  b) Multi-level hierarchy
  c) Long short term memory (LSTM)
  d) Faster hardware

  e) ReLU
  f) Batch normalization

2.12 Deep Learning with Jupyter

- Jupyter is an open source project. The Jupyter console is a terminal frontend for kernels using the Jupyter protocol. The console can be installed with: pip install jupyter-console
- Jupyter is a web application perfect for this task. Jupyter works with Notebooks, documents that mix rich text including beautifully rendered math formulas, blocks of code and code output, including graphics.
- Jupyter main features are :
  1. inline code execution
  2. easy idea structuring
  3. nice displays of pictures and dataframe
- Jupyter Notebook is an open source web interface that enables to include text, video, audio, images.
- The main difference between Jupyter console and Jupyter notebooks is that the console functions in interactive mode. Whenever you type a line of code, it is immediately executed and you can see the results.
- If you want to write medium-length pieces of code, do a deep exploration of a dataset to tell a story, the notebook is better. If you want to test out code you're writing, or run quick commands, the console is better.
- A kernel is a program that runs and introspects the user's code: it provides computation and communication with the frontend interfaces, such as notebooks.
- The Jupyter Notebook Application has three main kernels : the IPython, IRkernel and IJulia kernels.
- Anaconda is the most widely used Python distribution for data science and comes pre-loaded with all the most popular libraries and tools.
- There are quite a number of packages that will help you to deploy your notebooks and that are part of the Jupyter ecosystem.
  1. docker-stacks will come in handy when user need stacks of Jupyter applications and kernels as Docker containers.
  2. ipywidgets provides interactive HTML and JavaScript widgets for the Jupyter architecture that combine front-end controls coupled to a Jupyter kernel.

3. jupyter-drive allows IPython to use Google Drive for file management.
4. jupyter-sphinx-theme to add a Jupyter Sphinx theme to your notebook. It will make it easier to create intelligent and beautiful documentation.
5. kernel_gateway is a web server that supports different mechanisms for spawning and communicating with Jupyter kernels. Look here to see some use cases in which this package can be come in handy.

- To create a new notebook, click on the New button, and select Notebook. A new browser tab opens and shows the Notebook interface as follows :

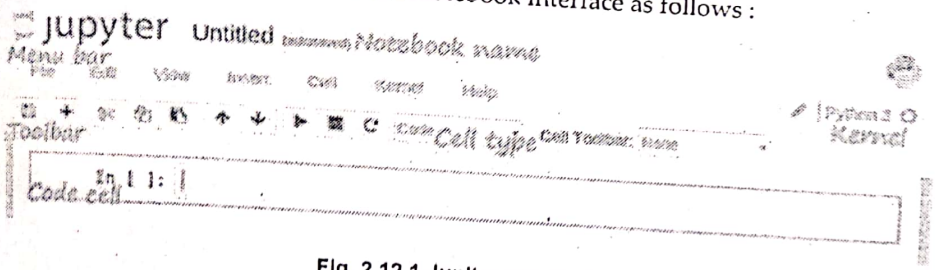


Fig. 2.12.1 Jupiter notebook

- Notebook consists of two cells : Markdown cells and code cell.
- Markdown cell is used for text and code cell for executing code by the kernel.
- Code cell contains various section like prompt number, input area, widget area and output area.
- Notebook model interface support edit mode and command mode.
- Try typing something like print("Hello World") into the cell. To run the code in the cell and see the output, click the Run button on the toolbar, or type Shift-Enter :

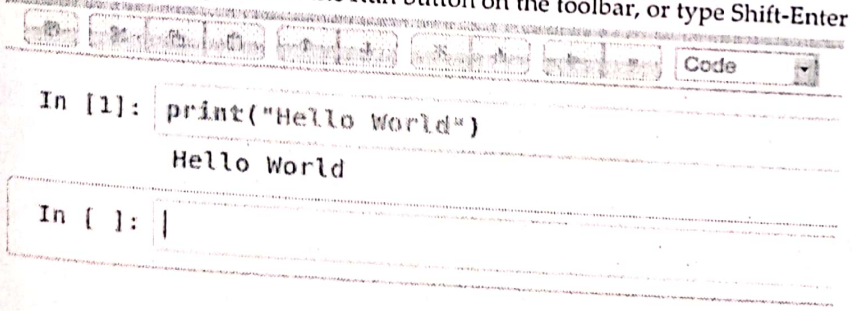


Fig. 2.12.2

- A Jupyter notebook is always organized as a sequence of so called 'cells' with each cell either containing some code or rich text created using the Markdown notation approach.

- When you click into the corresponding text field to add or modify the content of the cell, the bar color will change to green indicating that you are now in 'Edit mode'. Clicking anywhere outside of the text area of a cell will change back to 'Command mode'.

2.12.1 Magic Function

- Jupyter provides a number of so-called magic commands that can be used in code cells to simplify common tasks. Magic commands are interpreted by Jupyter and, for instance, transformed into Python code before the content is passed on to the kernel for execution
- Jupyter has a set of predefined 'magic functions' that you can call with a command line style syntax. There are two kinds of magics, line-oriented and cell-oriented.
- Line magics are prefixed with "%" and work like a command typed in your terminal. Line magics can be used as expression and their return value can be assigned to variable.
- Example of line magics are %autocall, %cd, %automagic, %dhist, %edit, %matplotlib and %env.
- Cell magics work on a block of code, not on a single line. They are prefixed with "%%". To close a block of code, when you are inside a cell magic function, hit Enter twice.

2.12.2 Using Jupyter Notebook

- The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.
- Jupyter Notebook is often used for exploratory data analysis and visualization. The Jupyter Notebook has several menus that you can use to interact with your Notebook. Here is a list of some current menus:

  a) Edit b) View C) Insert D) Cell E) Kernel f) Widgets G) Help H) File
- File menu is used for creating new notebook and opening existing one.
- Edit menu perform operation like cut, copy, and paste cells.
- View menu is useful for toggling the visibility of the header and toolbar.
- Insert menu is for inserting cells above or below the currently selected cell.
- Cell menu allows you to run one cell, a group of cells, or all the cells.
- The Widgets menu is for saving and clearing widget state.
- The Kernel cell is for working with the kernel that is running in the background

2.12.3 Restarting the Kernel

- Juypter has a autocomplete feature that helps with the code writing and can save a lot of typing : While editing code in a code cell, you can press the TAB key and Jupyter will either automatically complement the name or keyword you are writing or provide you with a dropdown list of choices that you can pick from.
- Fig. 2.12.3 shows restarting the kernel.

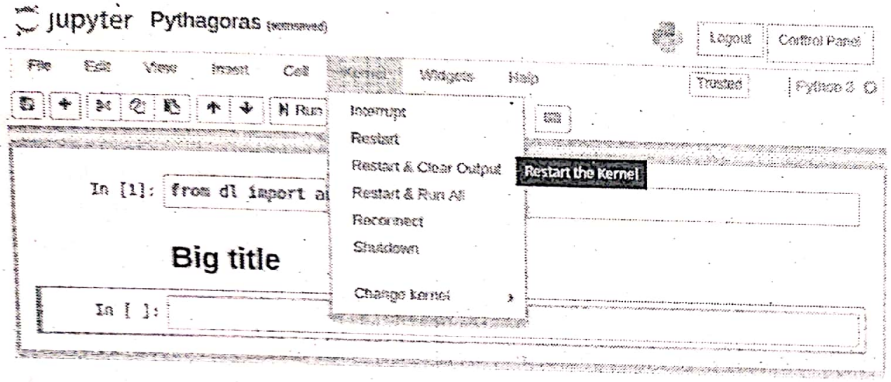


Fig. 2.12.3 Restarting the kernel

- Save your notebook locally to store your current progress
- In the notebook toolbar, click Kernel, then Restart
- Try testing your kernel by running a print statement in one of your notebook cells. If this is successful, you can continue to save and proceed with your work.
- If your notebook kernel is still timed out, try closing your browser and relaunching the notebook. When the notebook reopens, you will need to do "Cell -> Run All" or "Cell -> Run All Above" to regenerate the execution state.

2.12.4 Restoring a Checkpoint

- In Jupyter Notebook Workspaces, it's a good idea to save your checkpoints before you close the tab.
- To restore a checkpoint, choose the entry found on the File → Revert to Checkpoint menu.
- This menu makes it appear that you can have more than one checkpoint file, but the menu never has more than one entry in it.
- Fig. 2.12.4 shows menu for restoring a checkpoint.

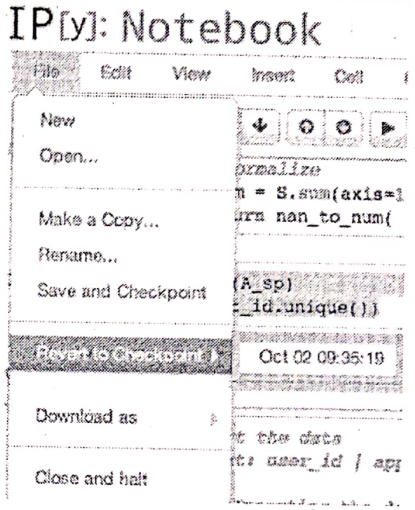


Fig. 2.12.4 Restoring a checkpoint

2.12.5 Performing Multimedia and Graphic Integration

- IPython Notebook support both a coding platform and a presentation platform.

Embedding plots and other images

- When you run %matplotlib inline, any plots you create appear as part of the document. Use %matplotlib inline - this only draws the images, not interactive / zoom-able but it works well.
- Following Fig. 2.12.5 shows example of embedding plot.

```
In [3]: 1 %matplotlib inline
        2 import matplotlib.pyplot as plt
        3 plt.plot([1.0,0],[0,1.0], linewidth=2, label='random diagonal')
        4 plt.show()
```

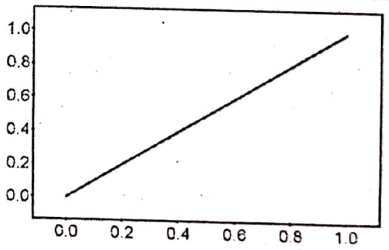


Fig. 2.12.5 Example of %matplotlib inline

- The display function is a general purpose tool for displaying different representations of objects. Think of it as print for these rich representations.

```
In [1]: from IPython.display import display
```

A few points :

a) Calling display on an object will send all possible representations to the Notebook.

b) These representations are stored in the Notebook document.

c) In general the Notebook will use the richest available representation.

- To work with images, use the Image class.

```
In [3]: from IPython.display import Image
In [4]: i = Image(url='img/python-logo.gif')
```

2.13 Deep Learning with Colab

- Colab (Colaboratory) is a product from Google Research, which is technically a hosted Jupyter Notebook service that requires no setup to use. It is cloud-based framework. With Google's Colab, one can access its GPUs and TPUs almost for free.
- Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning and data analysis.
- Colab has several useful features that make it helpful for the data science community. The following are some of the advantages :

a) Pre-installed data science libraries

b) Easy sharing and collaboration

c) Seamless integration with GitHub

d) Working with data from various sources

e) Automatic storage and version control

f) Access to hardware accelerators such as GPUs and TPUs

□□□

Notes