Quantum Series®
ENGINEERING

QUANTUM Series

Semester - 8 Computer Science & IT

Deep Learning

Session
2019-20
Even Semester

- Topic-wise coverage of entire syllabus in Question-Answer form.
- Short Questions (2 Marks)

QUANTUM SERIES

For
B.Tech Students of Fourth Year
of All Engineering Colleges Affiliated to
Dr. A.P.J. Abdul Kalam Technical University,
Uttar Pradesh, Lucknow
(Formerly Uttar Pradesh Technical University)

Deep Learning

By

Chetan Singhal

Kanika Dhama

Quantum
— Page —

QUANTUM PAGE PVT. LTD.
Ghaziabad ■ New Delhi

PUBLISHED BY : **Apram Singh**
Quantum Publications®
(A Unit of Quantum Page Pvt. Ltd.)
Plot No. 59/2/7, Site - 4, Industrial Area,
Sahibabad, Ghaziabad-201 010

Phone : 0120 - 4160479

Email : pagequantum@gmail.com **Website: www.quantumpage.co.in**

Delhi Office : 1/6590, East Rohtas Nagar, Shahdara, Delhi-110032

Information contained in this work is derived from sources believed to be reliable. Every effort has been made to ensure accuracy, however neither the publisher nor the authors guarantee the accuracy or completeness of any information published herein, and neither the publisher nor the authors shall be responsible for any errors, omissions, or damages arising out of use of this information.

Deep Learning (CS/IT : Sem-8)
1st Edition : *2019-20*

Price: Rs. 60/- only

Printed Version **:** e-Book.

CONTENTS

RCS 086 : DEEP LEARNING

Quantum Series®
ENGINEERING

QUANTUM *Series*

Related titles in Quantum Series

For Semester - 8
(Computer Science & IT)

Open Elective-2

- Digital and Social Media Marketing
- Entrepreneurship Development
- Machine Learning
- Operations Research
- Renewable Energy Resources
- Values, Relationship & Ethical Human Conduct

Department Elective-5

- Machine Learning
- Image Processing
- Parallel and Distributed Computing

Department Elective-6

- Deep Learning
- Data Compression

A comprehensive book to get the big picture without spending hours over lengthy text books.

Quantum Series is the complete one-stop solution for engineering student looking for a simple yet effective guidance system for core engineering subject. Based on the needs of students and catering to the requirements of the syllabi, this series uniquely addresses the way in which concepts are tested through university examinations. The easy to comprehend question answer form adhered to by the books in this series is suitable and recommended for student. The students are able to effortlessly grasp the concepts and ideas discussed in their course books with the help of this series. The solved question papers of previous years act as a additional advantage for students to comprehend the paper pattern, and thus anticipate and prepare for examinations accordingly.

The coherent manner in which the books in this series present new ideas and concepts to students makes this series play an essential role in the preparation for university examinations. The detailed and comprehensive discussions, easy to understand examples, objective questions and ample exercises, all aid the students to understand everything in an all-inclusive manner.

- Topic-wise coverage in Question-Answer form.
- Clears course fundamentals.
- Includes solved University Questions.
- The perfect assistance for scoring good marks.
- Good for brush up before exams.
- Ideal for self-study.

Quantum Page™

Quantum Publications®
(A Unit of Quantum Page Pvt. Ltd.)
Plot No. 59/2/7, Site-4, Industrial Area, Sahibabad,
Ghaziabad, 201010, (U.P.) Phone: 0120-4160479
E-mail: pagequantum@gmail.com Web: www.quantumpage.co.in
Find us on: **facebook.com/quantumseriesofficial**

1 UNIT

Introduction to Machine Learning

CONTENTS

PART-1

Introduction to Machine Learning, Linear Models (SVMs and Perceptrons, Logistic Regression).

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.1. Define the term Machine Learning.

Answer

1. Machine learning is an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
2. Machine learning focuses on the development of computer programs that can access data.
3. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.
4. Machine learning enables analysis of massive quantities of data.
5. It generally delivers faster and more accurate results in order to identify profitable opportunities or dangerous risks.
6. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

Que 1.2. What are the advantages and disadvantages of machine learning ?

Answer

Advantages of machine learning are :

1. **Easily identifies trends and patterns :**
   a. Machine learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans.
   b. For an e-commerce website like Flipkart, it serves to understand the browsing behaviours and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them.
   c. It uses the results to reveal relevant advertisements to them.
2. **No human intervention needed (automation) :** Machine learning does not require physical force *i.e.*, no human intervention is needed.

3. **Continuous improvement :**
   a. ML algorithms gain experience, they keep improving in accuracy and efficiency.
   b. As the amount of data keeps growing, algorithms learn to make accurate predictions faster.
4. **Handling multi-dimensional and multi-variety data :**
   a. Machine learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

Disadvantages of machine learning are :

1. **Data acquisition :**
   a. Machine learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality.
2. **Time and resources :**
   a. ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy.
   b. It also needs massive resources to function.
3. **Interpretation of results :**
   a. To accurately interpret results generated by the algorithms. We must carefully choose the algorithms for our purpose.
4. **High error-susceptibility :**
   a. Machine learning is autonomous but highly susceptible to errors.
   b. It takes time to recognize the source of the issue, and even longer to correct it.

Que 1.3. **Explain the components of machine learning system.**

Answer

Components of machine learning system are :

1. **Sensing :**
   a. It uses transducer such as camera or microphone for input.
   b. PR (Pattern Recognition) system depends on the bandwidth, resolution, sensitivity, distortion, etc., of the transducer.
2. **Segmentation :** Patterns should be well separated and should not overlap.
3. **Feature extraction :**
   a. It is used for distinguishing features.
   b. This process extracts invariant features with respect to translation, rotation and scale.

4. **Classification :**
   a. It use a feature vector provide by a feature extractor to assign the object to a category.
   b. It is not always possible to determine the values of all the features.
5. **Post processing :**
   a. Post processor uses the output of the classifier to decide on the recommended action.

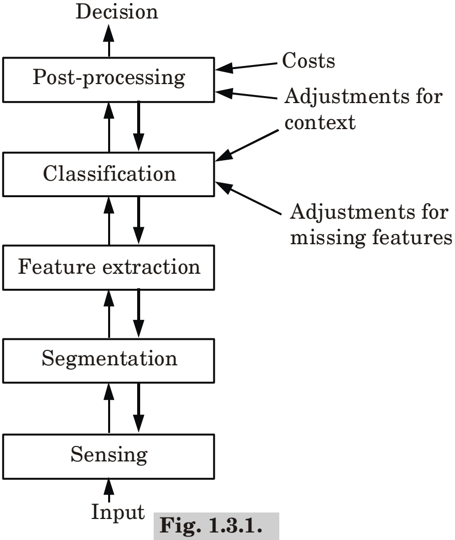


Fig. 1.3.1.

Que 1.4. **What are the classes of problem in machine learning ?**

Answer

Common classes of problem in machine learning :

1. **Classification :**
   a. In classification data is labelled *i.e.*, it is assigned a class, for example, spam/non-spam or fraud/non-fraud.
   b. The decision being modelled is to assign labels to new unlabelled pieces of data.
   c. This can be thought of as a discrimination problem, modelling the differences or similarities between groups.
2. **Regression :**
   a. Regression data is labelled with a real value rather than a label.
   b. The decision being modelled is what value to predict for new unpredicted data.

3. **Clustering :**
   a. In clustering data is not labelled, but can be divided into groups based on similarity and other measures of natural structure in the data.
   b. For example, organising pictures by faces without names, where the human user has to assign names to groups, like iPhoto on the Mac.
4. **Rule extraction:**
   a. In rule extraction, data is used as the basis for the extraction of propositional rules.
   b. These rules discover statistically supportable relationships between attributes in the data.

Que 1.5. **Briefly explain the issues related with machine learning.**

Answer

Issues related with machine learning are :

1. **Data quality :**
   a. It is essential to have good quality data to produce quality ML algorithms and models.
   b. To get high-quality data, we must implement data evaluation, integration, exploration, and governance techniques prior to developing ML models.
   c. Accuracy of ML is driven by the quality of the data.
2. **Transparency :**
   a. It is difficult to make definitive statements on how well a model is going to generalize in new environments.
3. **Manpower :**
   a. Manpower means having data and being able to use it. This does not introduce bias into the model.
   b. There should be enough skill sets in the organization for software development and data collection.
4. **Other :**
   a. The most common issue with ML is people using it where it does not belong.
   b. Every time there is some new innovation in ML, we are trying to use it where it is not necessary.
   c. This used to happen a lot with deep learning and neural networks.
   d. Traceability and reproduction of results are two main issues.

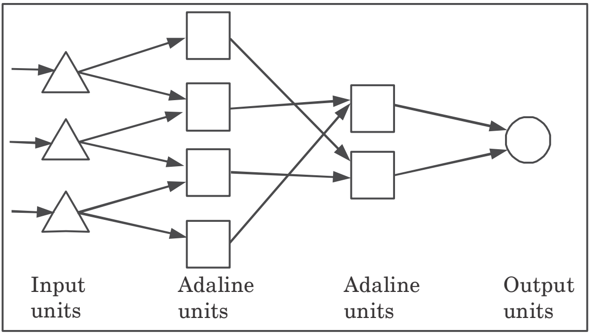
Que 1.6. **Define linear model in Machine Learning. Explain adaline network with its architecture.**

Answer

1. Linear model is defined as the model which is specified as a linear combination of features.
2. Based on training data, the learning process computes one weight for each feature to form a model that can predict or estimate the target value.

Adaline network :

1. ADALINE is an Adaptive Linear Neuron network with a single linear unit. The Adaline network is trained using the delta rule.
2. It receives input from several units and bias unit.
3. An Adaline model consists of trainable weights. The inputs are of two values (+ 1 or – 1) and the weights have signs (positive or negative).
4. Initially random weights are assigned. The net input calculated is applied to a quantizer transfer function (activation function) that restores the output to + 1 or – 1.
5. The Adaline model compares the actual output with the target output and with the bias units and then adjusts all the weights.



Que 1.7. **Explain SVMs linear model.**

Answer

1. A promising approach, which brings together the advantages of linear and non-linear models, follows the theory of Support Vector Machines (SVM).
2. In the case of two linearly separable classes, it is easy to find a dividing hyper plane, for example with the perceptron learning rule.
3. However, there are usually infinitely many such planes. We are looking for a plane which has the largest minimum distance to both classes. This plane is usually uniquely defined by a few points in the border area.

These points are called support vectors, all having the same distance to the dividing line.

4. To find the support vectors, there is an efficient optimizing algorithm. Optimal dividing hyper plane is determined by a few parameters, namely by the support vectors.
5. Support vector machines, apply this algorithm to non-linearly separable problems in a two-step process :
   a. In the first step, a non-linear transformation is applied to the data, with the property that the transformed data is linearly separable.
   b. In the second step, the support vectors are then determined in the transformed space.
6. It is always possible to make the classes linearly separable by transforming the vector space, as long as the data contains no contradictions.
7. Such a separation can be reached, for example by introducing a new $(n + 1)$th dimension and the definition,

$$x_{n+1} = \begin{Bmatrix} 1 & if\ x \in class\ 1 \\ 0 & if\ x \in class\ 2 \end{Bmatrix}$$

8. It can be shown that there are such generic transformations even for arbitrarily shaped class division boundaries in the original vector space. In the transformed space, the data are then linearly separable.
9. However, the number of dimensions of the new vector space grows exponentially with the number of dimensions of the original vector space.
10. However, the large number of new dimensions is not so problematic because, when using support vectors, the dividing plane, as mentioned above, is determined by only a few parameters.
11. The central non-linear transformation of the vector space is called the kernel, because of which support vector machines are also known as kernel methods.
12. The original SVM theory developed for classification has been extended and can now be used on regression problems also.

Que 1.8. **What is perceptron model ? Explain its working.**

Answer

1. The perceptron is the simplest form of a neural network used for classification of patterns said to be linearly separable.
2. It consists of a single neuron with adjustable synaptic weights and bias.
3. The perceptron build around a single neuron is limited for performing pattern classification with only two classes.

4. By expanding the output layer of perceptron to include more than one neuron, more than two classes can be classified.
5. Suppose, a perceptron have synaptic weights denoted by $w_1, w_2, w_3, \ldots.. w_m$.
6. The input applied to the perceptron are denoted by $x_1, x_2, \ldots\ldots x_m$.
7. The externally applied bias is denoted by b.

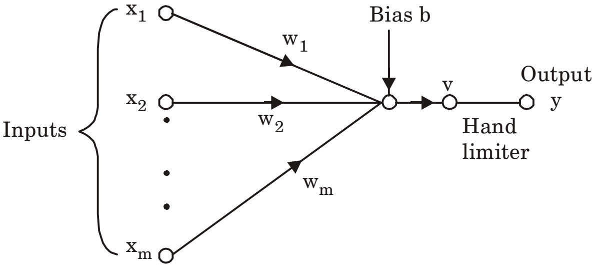


Fig. 1.8.1. Signal flow graph of the perceptron.

8. From the model, we find that the hard limiter input or induced local field of the neuron as

$$V = \sum_{i=1}^{m} w_i x_i + b$$

9. The goal of the perceptron is to correctly classify the set of externally applied input $x_1, x_2, \ldots\ldots x_m$ into one of two classes G_1 and G_2.
10. The decision rule for classification is that if output y is +1 then assign the point represented by input $x_1, x_2, \ldots\ldots. x_m$ to class G_1 else y is –1 then assign to class G_2.
11. In Fig. 1.8.2, if a point (x_1, x_2) lies below the boundary lines is assigned to class G_2 and above the line is assigned to class G_1. Decision boundary is calculated as :

$w_1x_1 + w_2x_2 + b = 0$

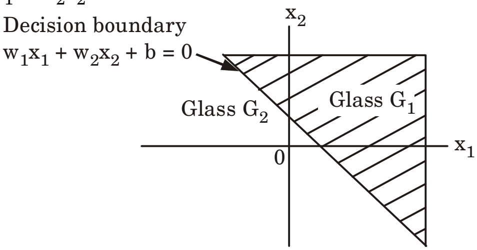


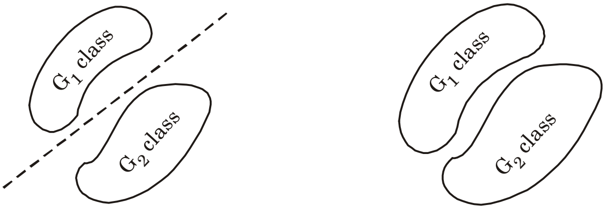
Fig. 1.8.2.

12. There are two decision regions separated by a hyperplane defined as :

$$\sum_{i=1}^{m} w_i x_i + b = 0$$

The synaptic weights $w_1, w_2, \ldots\ldots\ldots w_m$ of the perceptron can be adapted on an iteration by iteration basis.

13. For the adaption, an error-correction rule known as perceptron convergence algorithm is used.
14. For a perceptron to function properly, the two classes G_1 and G_2 must be linearly separable.
15. Linearly separable means, the pattern or set of inputs to be classified must be separated by a straight line.
16. Generalizing, a set of points in n-dimensional space are linearly separable if there is a hyperplane of $(n - 1)$ dimensions that separates the sets.



(*a*) A pair of linearly separable patterns

(*b*) A pair of non-linearly separable patterns

Fig. 1.8.3.

Que 1.9. **Write short note on logistic regression.**

Answer

1. Logistic regression is a supervised classification algorithm. It is based on maximum likelihood estimation.
2. In a classification problem, the target variable (output) y, can take only discrete values for given set of features (or inputs) x.
3. Logistic regression assumes the binomial distribution of the dependent variable. In logistic regression, we predict the value by 1 or 0.
4. Logistic regression builds a regression model to predict the probability that a given data entry belongs to the category numbered as 1. As linear regression models the data using the linear function, logistic regression models the data using the sigmoid function as :

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

5. Activation function is used to convert a linear regression equation to the logistic regression equation.

6. Logistic regression estimates the odds outcome of the dependent variable given a set of quantitative or categorical independent variables.
7. Logistic regression becomes a classification technique only when a decision threshold is used.
8. Any change in the coefficient leads to a change in both the direction and the steepness of the logistic function. It means positive slopes result in an S-shaped curve and negative slopes result in a Z-shaped curve.
9. Logistic regression is used to calculate the probability of an event.

PART-2

Introduction to Neural Networks : What a Shallow Network Computes.

Questions-Answers

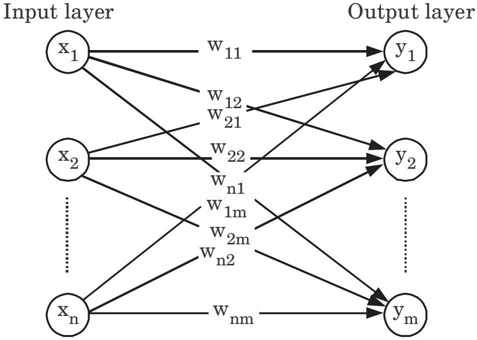
Long Answer Type and Medium Answer Type Questions

Que 1.10. **Explain different types of neuron connection with architecture.**

Answer

Different types of neuron connection are :

1. **Single-layer feed forward network :**



a. In this type of network, we have only two layers *i.e.*, input layer and output layer but input layer does not count because no computation is performed in this layer.

b. Output layer is formed when different weights are applied on input nodes and the cumulative effect per node is taken.

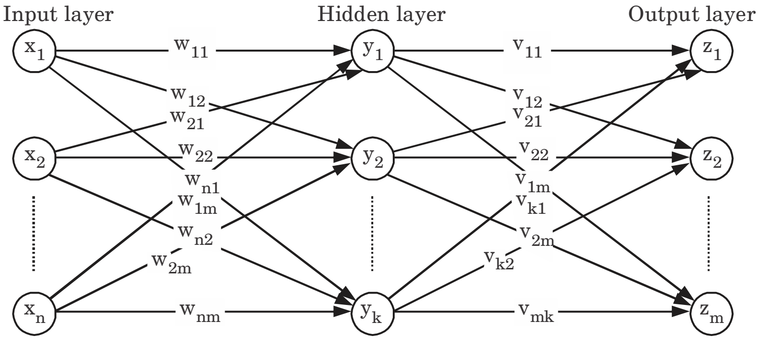
c. After this the neurons collectively give the output layer to compute the output signals.

2. **Multilayer feed forward network :**

a. This layer has hidden layer which is internal to the network and has no direct contact with the external layer.

b. Existence of one or more hidden layers enables the network to be computationally stronger.

c. There are no feedback connections in which outputs of the model are fed back into itself.



3. **Single node with its own feedback :**

a. When outputs can be directed back as inputs to the same layer or preceding layer nodes, then it results in feedback networks.

b. Recurrent networks are feedback networks with closed loop. Fig. 1.10.1 shows a single recurrent network having single neuron with feedback to itself.

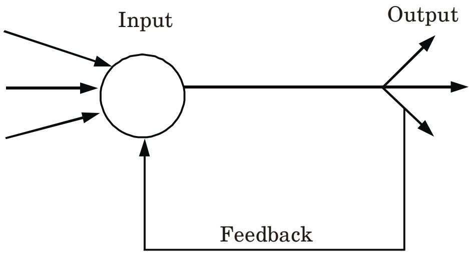


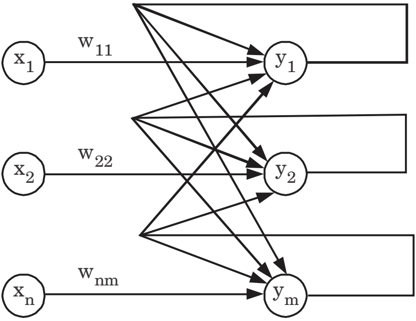
Fig. 1.10.1.

4. **Single-layer recurrent network :**

   a. This network is single layer network with feedback connection in which processing element's output can be directed back to itself or to other processing element or both.

   b. Recurrent neural network is a class of artificial neural network where connections between nodes form a directed graph along a sequence.

   c. This allows it to exhibit dynamic temporal behaviour for a time sequence. Unlike feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.

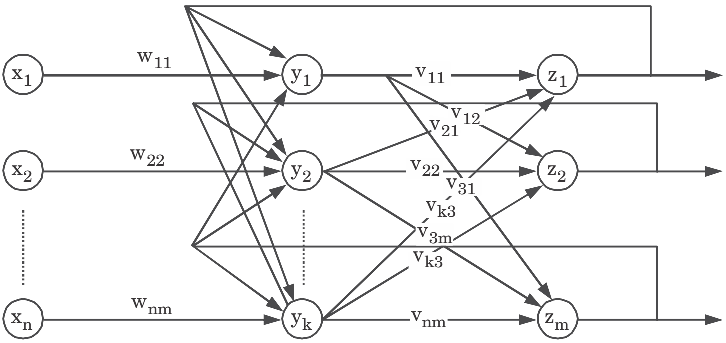


5. **Multilayer recurrent network :**

   a. In this type of network, processing element output can be directed to the processing element in the same layer and in the preceding layer forming a multilayer recurrent network.

   b. They perform the same task for every element of a sequence, with the output being depended on the previous computations. Inputs are not needed at each time step.

   c. The main feature of a multilayer recurrent neural network is its hidden state, which captures information about a sequence.



Que 1.11. Explain single layer neural network.

Answer

1. A single-layer neural network represents the most simple form of neural network, in which there is only one layer of input nodes that send weighted inputs to a subsequent layer of receiving nodes, or in some cases, one receiving node.
2. This single-layer design was part of the foundation for systems which have now become much more complex.
3. Single-layer neural networks can also be thought of as part of a class of feedforward neural networks, where information only travels in one direction, through the inputs, to the output.
4. Adaline network is an example of single layer neural network.

Adaline network : Refer Q. 1.6, Page 1–5M,Unit-1.

Que 1.12. Explain multilayer perceptron with its architecture and characteristics.

Answer

Multilayer perceptron :

1. The perceptrons which are arranged in layers are called multilayer perceptron. This model has three layers : an input layer, output layer and hidden layer.
2. For the perceptrons in the input layer, the linear transfer function used and for the perceptron in the hidden layer and output layer, the sigmoidal or squashed-S function is used.
3. The input signal propagates through the network in a forward direction.
4. On a layer by layer basis, in the multilayer perceptron bias $b(n)$ is treated as a synaptic weight driven by fixed input equal to +1.

   $$x(n) = [+1, x_1(n), x_2(n), \ldots\ldots\ldots x_m(n)]^T$$

   where n denotes the iteration step in applying the algorithm.

   Correspondingly, we define the weight vector as :

   $$w(n) = [b(n), w_1(n), w_2(n)\ldots\ldots\ldots\ldots, w_m(n)]^T$$
5. Accordingly, the linear combiner output is written in the compact form :

   $$V(n) = \sum_{i=0}^{m} w_i(n)x_i(n) = w^T(n) \times x(n)$$

The algorithm for adapting the weight vector is stated as :

1. If the nth number of input set $x(n)$, is correctly classified into linearly separable classes, by the weight vector $w(n)$ (that is output is correct) then no adjustment of weights are done.

$$w(n+1) = w(n)$$

if $w^T x(n) > 0$ and $x(n)$ belongs to class G_1.

$$w(n+1) = w(n)$$

if $w^T x(n) \leq 0$ and $x(n)$ belongs to class G_2.

2. Otherwise, the weight vector of the perceptron is updated in accordance with the rule.

Architecture of multilayer perceptron :

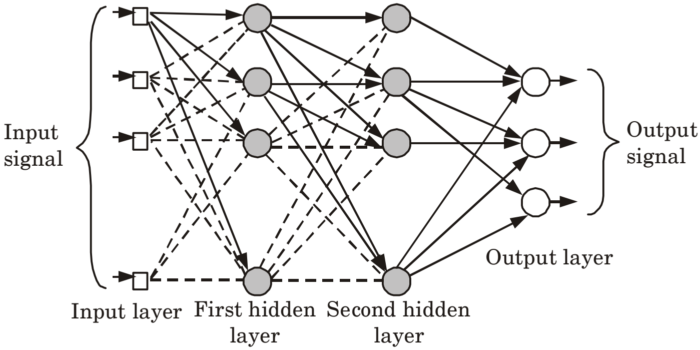


Fig. 1.12.1.

1. Fig. 1.12.1 shows architectural graph of multilayer perceptron with two hidden layer and an output layer.
2. Signal flow through the network progresses in a forward direction, from the left to right and on a layer-by-layer basis.
3. Two kinds of signals are identified in this network :
   a. **Functional signals :** Functional signal is an input signal and propagates forward and emerges at the output end of the network as an output signal.
   b. **Error signals :** Error signal originates at an output neuron and propagates backward through the network.
4. Multilayer perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with highly popular algorithm known as the error backpropagation algorithm.

Characteristics of multilayer perceptron :

1. In this model, each neuron in the network includes a non-linear activation function (non-linearity is smooth). Most commonly used non-linear function is defined by :

$$y_j = \frac{1}{1+\exp(-v_j)}$$

where v_j is the induced local field (*i.e.*, the sum of all weights and bias) and y is the output of neuron j.

2. The network contains hidden neurons that are not a part of input or output of the network. Hidden layer of neurons enabled network to learn complex tasks.
3. The network exhibits a high degree of connectivity.

Que 1.13. **Explain the statement : What a shallow network computes ?**

Answer

1. Shallow networks are the neural networks with less depth *i.e.*, less number of hidden layers.
2. These neural networks have one hidden layer and an output layer.
3. Shallow neural networks is a term used to describe neural network that usually have only one hidden layer as opposed to deep neural network which have several hidden layers.
4. Fig. 1.13.1, below shows a shallow neural network with single hidden layer, single input layer and single output layer :

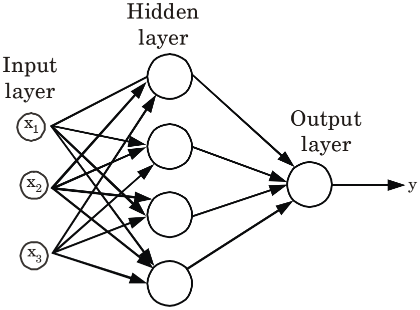


Fig. 1.13.1.

5. The neurons present in the hidden layer of our shallow neural network compute the following :

$$z_1^{[1]} = w_1^{[1]T}\ x + b_1^{[1]},\ a_1^{[1]} = \sigma\left(z_1^{[1]}\right) \quad ...(1.13.1)$$

$$z_2^{[1]} = w_2^{[1]T}\ x + b_2^{[1]},\ a_2^{[1]} = \sigma\left(z_2^{[1]}\right) \quad ...(1.13.2)$$

$$z_3^{[1]} = w_3^{[1]T}\ x + b_3^{[1]},\ a_3^{[1]} = \sigma\left(z_3^{[1]}\right) \quad ...(1.13.3)$$

$$z_4^{[1]} = w_4^{[1]T}\ x + b_4^{[1]},\ a_4^{[1]} = \sigma\left(z_4^{[1]}\right) \quad ...(1.13.4)$$

a. The superscript number [*i*] denotes the layer number and the subscript number *j* denotes the neuron number in a particular layer.

b. x is the input vector consisting of three features.

c. $W_j^{[i]}$ is the weight associated with neuron j present in the layer i.

d. $b_j^{[i]}$ is the bias associated with neuron j present in the layer i.

e. $Z_j^{[i]}$ is the intermediate output associated with neuron present in the layer i.

f. $a_j^{[i]}$ is the final output associated with neuron j present in the layer i.

6. Sigma is the sigmoid activation function. Mathematically, it is defined as :

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

As we can see, equations 1.13.1, 1.13.2, 1.13.3, and 1.13.4 are redundant. Therefore we will vectorize them as :

$$Z^{[1]} = X^{[1]T} X + b^{[1]} \quad ...(1.13.5)$$

$$A^{[1]} = \sigma(Z^{[1]}) \quad ...(1.13.6)$$

a. Equation (1.13.5) calculates the intermediate output $Z^{[1]}$ of the first hidden layer.

b. Equation (1.13.6) calculates the final output $A^{[1]}$ of the first hidden layer.

PART-3

Training a Network : Loss Functions, Backpropagation and Stochastic Gradient Descent, Neural Networks as Universal Function Approximates.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.14. Briefly explain training a network.

Answer

1. Once a network has been structured for a particular application, that network is ready to be trained.

2. To start this process the initial weights are chosen randomly. Then, the training begins.

3. There are two approaches to training :

a. **Supervised training :**

   i. In supervised training, both the inputs and the outputs are provided.

   ii. The network then processes the inputs and compares its resulting outputs against the desired outputs.

   iii. Errors are then propagated back through the system, causing the system to adjust the weights which controls the network.

   iv. This process occurs over and over as the weights are continually tweaked.

   v. The set of data which enables the training is called the "training set." During the training of a network the same set of data is processed many times as the connection weights are ever refined.

b. **Unsupervised (adaptive) training :**

   i. In unsupervised training, the network is provided with inputs but not with desired outputs.

   ii. The system itself must then decide what features it will use to group the input data. This is often referred to as self-organization or adaption.

   iii. This adaption to the environment is the promise which would enable science fiction types of robots to continually learn on their own as they encounter new situations and new environments.

Que 1.15. **Write short note on loss function.**

Answer

Loss function estimates how well particular algorithm models the provided data. Loss functions are classified into two classes based on the type of learning task as :

1. **Regression losses :**

   a. **Mean squared error (Quadratic Loss or L2 Loss) :** It is the average of the squared difference between predictions and actual observations.

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

   b. **Mean absolute error (L1 Loss) :** It is the average of sum of absolute difference between prediction and actual observation.

$$MAE = \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|^2}{n}$$

c. **Mean bias error :** It is less accurate but could conclude if the model has a positive bias or negative bias.

$$MBE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n}$$

d. **Huber loss (smooth mean absolute error) :** It is less sensitive to outliers in data than MSE and is also differentiable at 0. It is an absolute error, which becomes quadratic when the error is small.

$$\text{Loss} = \begin{cases} \frac{1}{2} * (x - y)^2 & \textit{if } (|x - y|) \leq \delta \\ \delta * |x - y| - \frac{1}{2} * \delta^2 & \text{Otherwise} \end{cases}$$

2. **Classification losses :**

a. **Cross entropy loss (negative log likelihood) :** It is the commonly used loss function for classification. Cross-entropy loss progress as the predicted probability diverges from actual label.

$$\text{Cross entropy loss} = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$

b. **Hinge loss (multi class SVM loss) :** Hinge loss is applied for maximum-margin classification, prominently for support vector machines. It is convex function used in convex optimizers.

$$\text{SVM Loss} = \sum_{j \neq y_i} \max(0, s_j - s_{yi} + 1)$$

Que 1.16. **Write a short note on backpropagation algorithm.**

Answer

1. Backpropagation is an algorithm used in the training of feedforward neural networks for supervised learning.
2. Backpropagation efficiently computes the gradient of the loss function with respect to the weights of the network for a single input-output example.
3. This makes it feasible to use gradient methods for training multi-layer networks, updating weights to minimize loss, we use gradient descent or variants such as stochastic gradient descent.
4. The backpropagation algorithm works by computing the gradient of the loss function with respect to each weight by the chain rule, iterating backwards one layer at a time from the last layer to avoid redundant calculations of intermediate terms in the chain rule; this is an example of dynamic programming.
5. The term backpropagation refers only to the algorithm for computing the gradient, but it is often used loosely to refer to the entire learning algorithm.

6. Backpropagation generalizes the gradient computation in the delta rule and is in turn generalized by automatic differentiation, where backpropagation is a special case of reverse accumulation (reverse mode).

Que 1.17. **How tuning parameters effect the backpropagation neural network ?**

Answer

Effect of tuning parameters of the backpropagation neural network :

1. **Momentum factor :**
   a. The momentum factor has a significant role in deciding the values of learning rate that will produce rapid learning.
   b. It determines the size of change in weights or biases.
   c. If momentum factor is zero, the smoothening is minimum and the entire weight adjustment comes from the newly calculated change.
   d. If momentum factor is one, new adjustment is ignored and previous one is repeated.
   e. Between 0 and 1 is a region where the weight adjustment is smoothened by an amount proportional to the momentum factor.
   f. The momentum factor effectively increases the speed of learning without leading to oscillations and filters out high frequency variations of the error surface in the weight space.
2. **Learning coefficient :**
   a. An formula to select learning coefficient has been :

$$h = \frac{1.5}{(N_1^{\,2} + N_2^{\,2} + ... + N_m^{\,2})}$$

   Where N_1 is the number of patterns of type 1 and m is the number of different pattern types.
   b. The small value of learning coefficient less than 0.2 produces slower but stable training.
   c. The largest value of learning coefficient *i.e.*, greater than 0.5, the weights are changed drastically but this may cause optimum combination of weights to be overshot resulting in oscillations about the optimum.
   d. The optimum value of learning rate is 0.6 which produce fast learning without leading to oscillations.
3. **Sigmoidal gain :**
   a. If sigmoidal function is selected, the input-output relationship of the neuron can be set as

$$O = \frac{1}{(1 + e^{-\lambda(1+\theta)})} \quad ...(1.17.1)$$

where λ is a scaling factor known as sigmoidal gain.

b. As the scaling factor increases, the input-output characteristic of the analog neuron approaches that of the two state neuron or the activation function approaches the (Satisifiability) function.

c. It also affects the backpropagation. To get graded output, as the sigmoidal gain factor is increased, learning rate and momentum factor have to be decreased in order to prevent oscillations.

4. **Threshold value :**

a. θ in equation (1.17.1) is called as threshold value or the bias or the noise factor.

b. A neuron fires or generates an output if the weighted sum of the input exceeds the threshold value.

c. One method is to simply assign a small value to it and not to change it during training.

d. The other method is to initially choose some random values and change them during training.

Que 1.18. Write short note on gradient descent.

Answer

1. Gradient descent is an optimization technique in machine learning and deep learning and it can be used with all the learning algorithms.
2. A gradient is the slope of a function, the degree of change of a parameter with the amount of change in another parameter.
3. Mathematically, it can be described as the partial derivatives of a set of parameters with respect to its inputs. The more the gradient, the steeper the slope.
4. Gradient descent is a convex function.
5. Gradient descent can be described as an iterative method which is used to find the values of the parameters of a function that minimizes the cost function as much as possible.
6. The parameters are initially defined a particular value and from that, Gradient descent is run in an iterative fashion to find the optimal values of the parameters, using calculus, to find the minimum possible value of the given cost function.

Que 1.19. Discuss selection of various parameters in Backpropagation Neural Network (BPN).

Answer

Selection of various parameters in BPN :

1. **Number of hidden nodes :**
   a. The guiding criterion is to select the minimum nodes in the first and third layer, so that the memory demand for storing the weights can be kept minimum.

   b. The number of separable regions in the input space M, is a function of the number of hidden nodes H in BPN and H = M – 1.

   c. When the number of hidden nodes is equal to the number of training patterns, the learning could be fastest.

   d. In such cases, BPN simply remembers training patterns losing all generalization capabilities.

   e. Hence, as far as generalization is concerned, the number of hidden nodes should be small compared to the number of training patterns with help of VCdim (Vapnik Chervonenkis dimension) probability theory.

   f. We can estimate the selection of number of hidden nodes for a given number of training patterns as number of weights which is equal to $I_1 * I_2 + I_2 * I_3$, where I_1 and I_3 denote input and output nodes and I_2 denote hidden nodes.

   g. Assume the training samples T to be greater than VCdim. Now if we accept the ratio 10 : 1

$$10 * T = \frac{I_2}{(I_1 + I_3)}$$

$$I_2 = \frac{10T}{(I_1 + I_3)}$$

   Which yields the value for I_2.

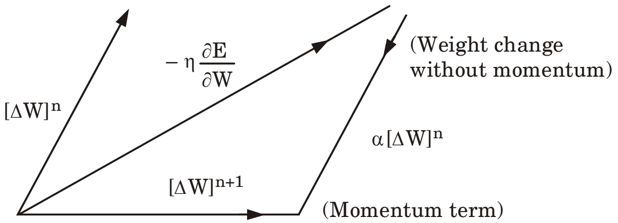
2. **Momentum coefficient α :**
   a. To reduce the training time we use the momentum factor because it enhances the training process.

   b. The influences of momentum on weight change is

$$[\Delta W]^{n+1} = -\eta \frac{\partial E}{\partial W} + \alpha[\Delta W]^{n}$$

   c. The momentum also overcomes the effect of local minima.

   d. The use of momentum term will carry a weight change process through one or local minima and get it into global minima.



$-\eta \frac{\partial E}{\partial W}$
(Weight change without momentum)
$[\Delta W]^n$
$\alpha [\Delta W]^n$
$[\Delta W]^{n+1}$
(Momentum term)

Fig. 1.19.1. Influence of momentum term on weight change.

3. **Sigmoidal gain λ :**
   a. When the weights become large and force the neuron to operate in a region where sigmoidal function is very flat, a better method of coping with network paralysis is to adjust the sigmoidal gain.
   b. By decreasing this scaling factor, we effectively spread out sigmoidal function on wide range so that training proceeds faster.
4. **Local minima :**
   a. One of the most practical solutions involves the introduction of a shock which changes all weights by specific or random amounts.
   b. If this fails, then the most practical solution is to rerandomize the weights and start the training all over.

Que 1.20. **Explain different types of gradient descent.**

Answer

Different types of gradient descent are :

1. **Batch gradient descent :**
   a. This is a type of gradient descent which processes all the training example for each iteration of gradient descent.
   b. When the number of training examples is large, then batch gradient descent is computationally very expensive. So, it is not preferred.
   c. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.
2. **Stochastic gradient descent :**
   a. This is a type of gradient descent which processes single training example per iteration.
   b. Hence, the parameters are being updated even after one iteration in which only a single example has been processed.
   c. Hence, this is faster than batch gradient descent. When the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be large.

3. **Mini-batch gradient descent :**
   a. This is a mixture of both stochastic and batch gradient descent.
   b. The training set is divided into multiple groups called batches.
   c. Each batch has a number of training samples in it.
   d. At a time, a single batch is passed through the network which computes the loss of every sample in the batch and uses their average to update the parameters of the neural network.

Que 1.21. **What are the advantages and disadvantages of stochastic gradient descent ?**

Answer

Advantages of stochastic gradient descent :

1. It is easier to fit into memory due to a single training sample being processed by the network.
2. It is computationally fast as only one sample is processed at a time.
3. For larger datasets it can converge faster as it causes updates to the parameters more frequently.
4. Due to frequent updates the steps taken towards the minima of the loss function have oscillations which can help getting out of local minimums of the loss function (in case the computed position turns out to be the local minimum).

Disadvantages of stochastic gradient descent :

1. Due to frequent updates the steps taken towards the minima are very noisy. This can often lead the gradient descent into other directions.
2. Also, due to noisy steps it may take longer to achieve convergence to the minima of the loss function.
3. Frequent updates are computationally expensive due to using all resources for processing one training sample at a time.
4. It loses the advantage of vectorized operations as it deals with only a single example at a time.

Que 1.22. **Explain neural networks as universal function approximation.**

Answer

1. Feedforward networks with hidden layers provide a universal approximation framework.
2. The universal approximation theorem states that a feedforward network with a linear output layer and atleast one hidden layer with any "squashing" activation function can approximate any Borel measurable function from one finite-dimensional space to another with any desired non-zero amount of error, provided that the network is given enough hidden units.

3. The derivatives of the feedforward network can also approximate the derivatives of the function.
4. The concept of Borel states that for any continuous function on a closed and bounded subset of R^n is Borel measurable and therefore may be approximated by a neural network.
5. The universal approximation theorem means that a large MLP will be able to represent function.
6. Even if the MLP is able to represent the function, learning can fail for two different reasons.
   a. Optimization algorithm used for training may not be able to find the value of the parameters that corresponds to the desired function.
   b. Training algorithm might choose the wrong function due to overfitting.
7. Feedforward networks provide a universal system for representing functions, in the sense that, given a function, there exists a feedforward network that approximates the function.
8. There is no universal procedure for examining a training set of specific examples and choosing a function that will generalize to points not in the training set.
9. The universal approximation theorem says that there exists a network large enough to achieve any degree of accuracy we desire, but the theorem does not say how large this network will be.
10. Scientists provide some bounds on the size of a single-layer network needed to approximate a broad class of functions. In the worse case, an exponential number of hidden units.
11. This is easiest to see in the binary case : the number of possible binary functions on vectors $v \in \{0,1\}^n$ is 2^{2^n} and selecting one such function requires 2^n bits, which will in general require $O(2^n)$ degrees of freedom.
12. A feedforward network with a single layer is sufficient to represent any function, but the layer may be infeasibly large and may fail to learn and generalize correctly.

☺☺☺

2

UNIT

Deep Networks

CONTENTS

PART- 1

History of Deep Learning, A Probabilistic Theory of Deep Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.1. **What do you understand by deep learning ?**

Answer

1. Deep learning is the subfield of artificial intelligence that focuses on creating large neural network models that are capable of making accurate data-driven decisions.
2. Deep learning is used where the data is complex and has large datasets.
3. Facebook uses deep learning to analyze text in online conversations. Google and Microsoft all use deep learning for image search and machine translation.
4. All modern smart phones have deep learning systems running on them. For example, deep learning is the standard technology for speech recognition, and also for face detection on digital cameras.
5. In the healthcare sector, deep learning is used to process medical images (X-rays, CT, and MRI scans) and diagnose health conditions.
6. Deep learning is also at the core of self-driving cars, where it is used for localization and mapping, motion planning and steering, and environment perception, as well as tracking driver state.

Que 2.2. **Explain the history of deep learning.**

Answer

1. **In 300 BC :** Aristotle introduce associationism, started the history of human's attempt to understand brain.
2. **In 1873 :** Alexander Bain introduce neural groupings as the earliest models of neural network.
3. **In 1913 :** MeCulloch and Pitts introduce MCP model, which is considered as the ancestor of artificial neural model.
4. **In 1919 :** Donald Hebb considered as the father of neural networks, introduced Hebbian Learning Rule, which lays the foundation of modern neural network.

5. **In 1958 :** Frank Rosenblatt introduce the first perception, which highly resembles modern perception.
6. **In 1974 :** Paul Werbos introduce backpropagation.
7. **In 1980 :** Tenvo Kohonen introduce self organizing map.
8. **In 1980 :** Kumihiko Fukushima introduce Neocognitron, which inspired convolutional neural network.
9. **In 1982 :** John Hopfield introduce Hopfield network.
10. **In 1985 :** Hilton and Sejnowski introduce Boltzmann machine.
11. **In 1986 :** Paul Smolensky introduce Harmonium, which is later known as restricted Boltzmann machine.
12. **In 1986 :** Michael I. Jordan defined and introduce recurrent neural network.
13. **In 1990 :** Yann LeCun introduce LeNet, showed the possibility of deep neural networks in practice.
14. **In 1997 :** Scluster and Paliwal introduce bidirectional recurrent neural network.
15. **In 2006 :** Geoffrey Hinton introduce deep belief networks, also introduced layer-wise pretraining technique, opened current deep learning era.
16. **In 2009 :** Salakhutdinov and Hinton introduce deep Boltzmann machines.
17. **In 2012 :** Geoffrey Hinton introduce Dropont, an efficient way of training neural networks.

Que 2.3. **Discuss the algorithm used for deep learning.**

Answer

Following are the algorithm used for deep learning :

1. **Feed forward neural networks :**
   a. A feed forward neural network is an artificial neural network wherein connections between the nodes do not form a cycle.
   b. Feedforward neural networks are used for supervised learning in cases where the data to be learned is neither sequential nor time-dependent.
2. **Radial basis function neural network :**
   a. A radial basis function network is an artificial neural network that uses radial basis functions as activation functions.
   b. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters.
3. **Multilayer perceptron :** Refer Q. 1.12, Page 1–13M, Unit-1.

4. **Unsupervised pre-trained network :**
   a. Unsupervised pre-training initializes a discriminative neural network from one which was trained using an unsupervised criterion, such as a deep belief network or a deep autoencoder.

   b. This method help with both the optimization and the overfitting issues :

      i. **Autoencoders :**

         a. An autoencoder is a type of artificial neural network used to learn efficient data coding in an unsupervised manner.

         b. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal noise.

      ii. **Deep belief networks :**

         a. A Deep Belief Network (DBN) is a generative graphical model composed of multiple layers of latent variables (hidden units), with connections between the layers but not between units within each layer.

      iii. **Generative Adversarial Networks (GANs) :**

         a. Generative Adversarial Networks (GANs) are a powerful class of neural networks that are used for unsupervised learning.

         b. GANs are basically made up of a system of two competing neural network models which compete with each other and are able to analyze, capture and copy the variations within a dataset.

5. **Convolutional Neural Networks (CNNs) :**
   a. ConvNets (CNNs) are the category of Neural Networks that have proven very effective in areas such as image recognition and classification.

   b. ConvNets have been successful in identifying faces, objects and traffic signs apart from powering vision in robots and self driving cars.

6. **Recurrent neural network :**
   a. A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence.

   b. This allows it to exhibit temporal dynamic behavior.

   c. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs.

7. **Recursive neural networks :**

   a. A recursive neural network is a kind of deep neural network created by applying the same set of weights recursively over a structured input, to produce a structured prediction over variable-size input structures, or a scalar prediction on it, by traversing a given structure in topological order.

Que 2.4. **Differentiate between machine learning and deep learning.**

Answer

S. No.	Machine learning	Deep learning
1.	Works on small amount of dataset for accuracy.	Works on large amount of dataset.
2.	Dependent on low-end machine.	Heavily dependent on high-end machine.
3.	Divides the tasks into sub-tasks, solves them individually and finally combine the results.	Solves problem end to end.
4.	Takes less time to train.	Takes more time to train.
5.	More time to test the data.	Less time to test the data.

Que 2.5. **What are the applications of deep learning ?**

Answer

1. **Automatic text generation :**

   a. Corpus of text is learned and from this model new text is generated, word-by-word, character-by-character.

   b. Then this model is capable of learning how to spell, punctuate, form sentences, or it may even capture the style.

2. **Healthcare :** Helps in diagnosing various diseases and treating it.

3. **Automatic machine translation :** Certain words, sentences or phrases in one language is transformed into another language.

4. **Image recognition :** Recognizes and identifies peoples and objects in images as well as to understand content and context. This area is already being used in gaming, retail, tourism, etc.

5. **Predicting earthquakes :** Teaches a computer to perform viscoelastic computations which are used in predicting earthquakes.

Que 2.6. Write short note on deep networks.

Answer

1. Deep Neural Networks (DNNs), also called convolutional networks, are composed of multiple levels of nonlinear operations, such as neural networks with many hidden layers.
2. Deep learning methods aim at learning feature hierarchies, where features at higher levels of the hierarchy are formed using the features at lower levels.
3. Deep learning networks are distinguished from single hidden layer neural networks by their depth *i.e.*, the number of node layers through which data must pass in a multistep process of pattern recognition.
4. Earlier versions of neural networks such as the first perceptrons were shallow, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as deep learning.
5. In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further we advance into the neural networks, the more complex the features our nodes can recognize, since they aggregate and recombine features from the previous layer.

Que 2.7. What do you understand by probabilistic theory of deep learning ?

Answer

1. Probability is the science of quantifying uncertain things.
2. Most of machine learning and deep learning systems utilize a lot of data to learn about patterns in the data.
3. Whenever data is utilized in a system rather than sole logic, uncertainty grows up and whenever uncertainty grows up, probability becomes relevant.
4. In deep learning, several models like bayesian models, probabilistic graphical models, hidden Markov models are used. They depend entirely on probability concepts.
5. Real world data is disordered. Since deep learning systems utilize real world data, they require a tool to handle the disorderness.
6. It is always practical to use a simple and uncertain system rather than a complex but certain and brittle one.
7. For example, in the Fig. 2.7.1, the input layer is a flattened vector of the size of the input image (28*28 = 784).

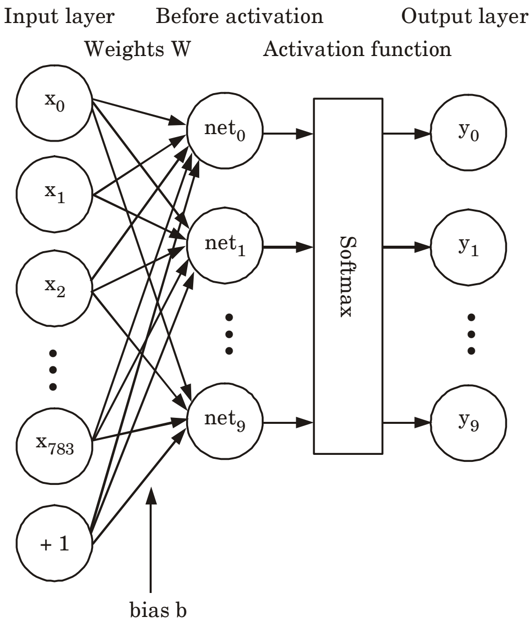


Fig. 2.7.1.

8. It is passed to a layer, where the input vector is multiplied by the weights and added with the bias vector.
9. This layer has 10 neurons. This is the implication that there are 10 digits. Then they go through a softmax activation function.
10. After this step they do not output the exact digit but a vector of length 10 with each element being a probability value for each digit.
11. We use argmax to get the index of the probability with the highest value in the output vector (which is the prediction).

PART-2

Backpropagation and Regularization, Batch Normalization.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.8. **Explain the backpropagation algorithm in detail.**

Answer

Refer Q. 1.16, Page 1–18M, Unit-1.

Que 2.9. **Define regularization.**

Answer

1. Regularization is the process used to reduce the complexity of the regression function without actually reducing the degree of the underlying polynomial function.
2. This technique is based on the fact that if the highest order terms in a polynomial equation have very small coefficients, then the function will approximately behave like a polynomial function of a smaller degree.
3. Regularization is done by adding a complexity term to the cost function which will give a higher cost as the complexity of the underlying polynomial function increases.

$$J(\theta) = \sum_{m} (\theta^T x - y)^2 + \lambda\theta^2$$

4. The formula is given in matrix form. The squared terms represent the squaring of each element of the matrix.
5. Regularised regressions are categorized on the basis of the complexity terms added to the cost function.

Que 2.10. **What is batch normalization ?**

Answer

1. Batch normalization is a technique for training every deep neural network that standardizes the inputs to a layer for each mini-batch.
2. This has the effect of stabilizing the learning process and reducing the number of training epochs required to train deep networks.
3. Batch normalization allows us to use much higher learning rates, which further increases the speed at which networks train.
4. Makes weights easier to initialise. Weight initialisation can be difficult, especially when creating deeper networks.
5. Batch normalization helps to reduce the sensitivity to the initial starting weights.
6. Batch normalization makes the input to each layer to have zero mean and unit variance.
7. Regularization reduces overfitting which leads to better test performance through better generalization.

8. We cannot use batch normalization on a recurrent neural network, as the statistics are computed per batch, this does not consider the recurrent part of the network.
9. Weights are shared in an RNN, and the activation response for each recurrent loop might have completely different statistical properties.

Que 2.11. **What are the advantages and disadvantages of batch normalization ?**

Answer

Advantages of batch normalization :

1. It reduces internal covariant shift.
2. It reduces the dependence of gradients on the scale of the parameters or their initial values.
3. Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
4. It allows use of saturating nonlinearities and higher learning rates.

Disadvantages of batch normalization :

1. Difficult to estimate mean and standard derivation of input during testing.
2. It cannot use batch size of one during training.
3. Computational overhead occurs during training.

PART-3

VC Dimension and Neural Networks, Deep Vs Shallow Networks Convolutional Networks Generative Adversarial Networks (GAN), Semi-supervised Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.12. **Explain VC dimension for neural networks.**

Answer

1. The VC dimension measures the capacity of a binary classifier.
2. The VC dimension is defined as being the largest possible value of m for which there exists a training set of m different x points that the classifier can label arbitrarily.

3. The VC dimension is a combinatorial characterization of the diversity of functions that can be computed by a given neural architecture.
4. The VC dimension can be regarded as a measure of the capacity or expressive power of a network.
5. VC dimension is the measure of model complexity (capacity) used in VC theory.
6. For linear estimators, the VC dimension is equivalent to the number of model parameters, but is hard to obtain for other types of estimators.
7. *A* subset S of the domain X is shattered by a class of functions or neural network N if every function $f : S \rightarrow \{0, 1\}$ can be computed on N.
8. The VC dimension of N is defined as the maximal size of a set $S \subseteq X$ that is shattered by N

   $$\text{dimvc}(N) = \max \{ | S | S \subseteq X \text{ is shattered by } N\} \quad ...(2.12.1)$$

   where $| S |$ denotes the cardinality of S.
9. For example, for a neural network with the relation $f(x, w, \theta) = \text{sgn}(w^T x + \theta)$, it can shatter at most any three points in X, thus its *VC* dimension is 3. This is shown in Fig. 2.12.1.

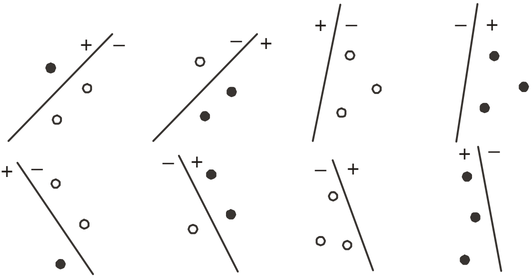


Fig. 2.12.1. Shatter any three points X into two classes.

10. The points are in general position, that is, they are linearly independent.
11. A hard-limiter function with threshold θ_0 is used as the activation function for binary neurons.
12. The basic function of the McCulloch-Pitts neuron has a linear relation applied by a threshold operation, hence called a Linear Threshold Gate (LTG).
13. A neural network with LTG has a VC dimension of $O(N_w \log N_w)$, where N_w is the number of weights in a network.
14. The VC dimension has been generalized for neural networks with real-valued output, and the VC dimension of various neural networks has been studied in.
15. The VC dimension can be used to estimate the number of training examples for a good generalization capability.

16. The Boolean VC dimension of a neural network N, written dimBVC(N), is defined as the VC dimension of the class of Boolean functions that is computed by N. The VC dimension is a property of a set of functions $\{f(\alpha)\}$, and can be defined for various classes of function f.
17. The VC dimension for the set of function $\{f(\alpha)\}$ is defined as the maximum number of training points that can be shattered by $\{f(\alpha)\}$.
18. If the VC dimension is d, then there exists atleast one set of d point that can be shattered, but in general it will not be true that every set of d points can be shattered.

Que 2.13. **Write short note on neural network and shallow neural network.**

Answer

Neural networks :

1. A neural network is a series of algorithms that endeavors to recognize relationships in a set of data through a process that mimics the way the human brain operates.
2. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature.
3. Neural networks can adapt to changing input. So the network generates the best possible result without needing to redesign the output criteria.
4. The concept of neural networks, which has its roots in artificial intelligence and in the development of trading systems.
5. A neural network contains layers of interconnected nodes.
6. Each node is a perceptron and is similar to a multiple linear regression.
7. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear.
8. Neural networks are used, with applications for financial operations, enterprise planning, trading, business analytics and product maintenance.
9. Neural networks have also gained widespread adoption in business applications such as forecasting and marketing research solutions, fraud detection and risk assessment.

Shallow neural network : Refer Q. 1.13, Page 1–15M, Unit-1.

Que 2.14. **Define convolutional networks.**

Answer

1. Convolutional networks also known as Convolutional Neural Networks (CNNs) are a specialized kind of neural network for processing data that has a known, grid-like topology.

2. Convolutional neural network indicates that the network employs a mathematical operation called convolution.
3. Convolution is a specialized kind of linear operation.
4. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in atleast one of their layers.
5. CNNs (ConvNets) are quite similar to regular neural networks.
6. They are still made up of neurons with weights that can be learned from data. Each neuron receives some inputs and performs a dot product.
7. They still have a loss function on the last fully connected layer.
8. They can still use a non-linearity function a regular neural network receives input data as a single vector and passes through a series of hidden layers.

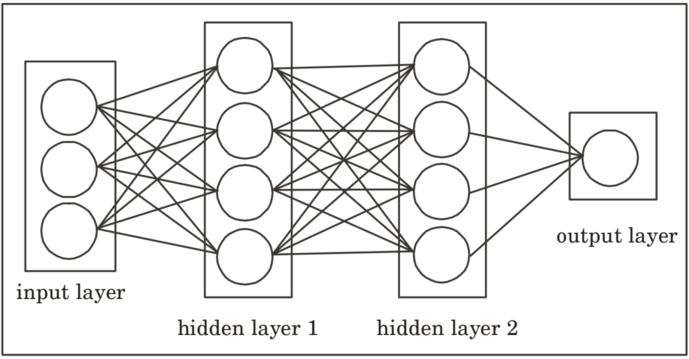


Fig. 2.14.1. A regular three-layer neural network.

9. Every hidden layer consists of neurons, wherein every neuron is fully connected to all the other neurons in the previous layer.
10. Within a single layer, each neuron is completely independent and they do not share any connections.
11. The fully connected layer (the output layer) contains class scores in the case of an image classification problem.

Que 2.15. **Discuss Generative Adversarial Network (GAN).**

Answer

1. Generative Adversarial Networks (GANs) are a powerful class of neural networks that are used for unsupervised learning.
2. GANs are made up of a system of two competing neural network models which compete with each other and are able to analyze, capture and copy the variations within a dataset.

3. Generative Adversarial Networks (GANs) can be broken down into three parts :
   a. **Generative :** To learn a generative model, which describe how data is generated in terms of a probabilistic model.
   b. **Adversarial :** The training of a model is done in an adversarial setting.
   c. **Networks :** Use deep neural networks as the Artificial Intelligence (AI) algorithms for training purpose.
4. In GANs, there is a generator and a discriminator. The Generator generates fake samples of data and tries to distract the discriminator.
5. The Discriminator tries to distinguish between the real and fake samples. The Generator and the Discriminator are both Neural Networks and they both run in competition with each other in the training phase as shown in Fig. 2.15.1.

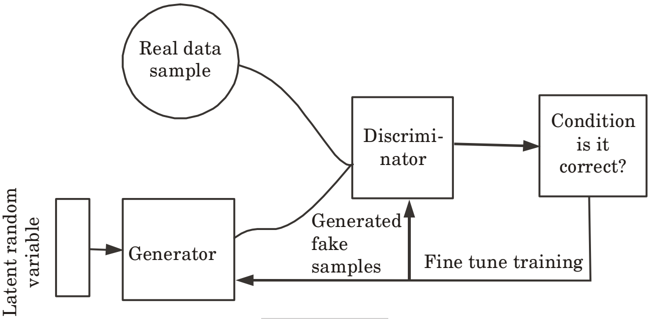


Fig. 2.15.1.

6. Here, the generative model captures the distribution of data and is trained in such a manner that it tries to maximize the probability of the Discriminator in making a mistake.
7. The Discriminator is based on a model that estimates the probability that the sample that it got is received from the training data and not from the Generator.
8. The GANs are formulated as a minimax game, where the Discriminator is trying to minimize its reward $V(D, G)$ and the Generator is trying to minimize the Discriminator's reward or in other words, maximize its loss.

Que 2.16. **What are the advantages and disadvantages of GAN ?**

Answer

Advantages of GAN :

1. Better modeling of data distribution (images sharper and clearer).

2. GANs can train any kind of generator network. Other frameworks require generator networks to have some specific form of functionality, such as the output layer being Gaussian.
3. There is no need to use the Markov chain to repeatedly sample, without inferring in the learning process, without complicated variational lower bounds, avoiding the difficulty of approximating the difficult probability of calculation.

Disadvantages of GAN :

1. Hard to train, unstable. Good synchronization is required between the generator and the discriminator.
2. Mode collapse issue. The learning process of GANs may have a missing pattern, the generator begins to degenerate, and the same sample points are always generated, and the learning cannot be continued.

Que 2.17. **Define semi-supervised learning.**

Answer

1. In semi-supervised learning, the algorithm is trained upon a combination of labeled and unlabeled data.
2. This combination will contain a very small amount of labeled data and a very large amount of unlabeled data.
3. The basic procedure involved is that the programmer will cluster similar data using an unsupervised learning algorithm and then use the existing labeled data to label the rest of the unlabeled data.
4. The typical use cases of such type of algorithm have a common property among them. The acquisition of unlabeled data is relatively cheap while labeling the data is very expensive.
5. A semi-supervised algorithm assumes the following about the data :
    a. **Continuity assumption :** The algorithm assumes that the points which are closer to each other are more likely to have the same output label.
    b. **Cluster assumption :** The data can be divided into discrete clusters and points in the same cluster are more likely to share an output label.
    c. **Manifold assumption :** The data lies approximately on a manifold of much lower dimension than the input space. This assumption allows the use of distances and densities which are defined on a manifold.

Que 2.18. **What are the applications of semi-supervised learning ?**

Answer

Applications of semi-supervised learning :

1. **Speech analysis :** Since labeling of audio files is a very intensive task, Semi-Supervised learning is a very natural approach to solve this problem.
2. **Internet content classification :**
   a. Labeling each webpage is an impractical and unfeasible process and thus uses semi-supervised learning algorithms.
   b. Even the Google search algorithm uses a variant of Semi-Supervised learning to rank the relevance of a webpage for a given query.
3. **Protein sequence classification :** Since DNA strands are typically very large in size, the rise of semi-supervised learning has been imminent in this field.

Que 2.19. **What are different types of semi-supervised learning algorithm ?**

Answer

Different types of semi-supervised learning algorithm are :

1. **Self training :**
   a. This is a wrapper algorithm and is the most commonly used technique. In self training, a classifier is trained on labeled data.
   b. Then, this classifier is used to classify all unlabeled items.
   c. The unlabeled items that are predicted with the highest confidence are added to the training set.
   d. Now the classifier is trained again on the training set and the above process is repeated.
2. **Generative models :**
   a. In this method, we assume the form of joint probability $p(x, y \mid \theta) = p(y \mid \theta)p(x \mid y, \theta)$ for semi-supervised learning.
   b. Parameters of joint probability are represented by $\theta \in \Theta$. Predictors f_0 use Bayes rule :

$$f_\theta(x) \equiv \underset{y}{\arg\max}\ p(y \mid x, \theta) = \underset{y}{\arg\max} \frac{p(x \mid y, \theta)}{\sum_y p(x \mid y, \theta)}$$

3. **Co-training :**
   a. The idea of co-training is to train two classifiers which then teach each other.
   b. It is a wrapper algorithm. There are two assumptions in co-training :

i. Data x can be split into two views $[x^{(1)}, x^{(2)}]$. Each view alone is enough to train a classifier, given enough labeled data.

ii. These two views are conditionally independent.

4. **Graph based methods :**

a. In this method, a graph is constructed.

b. The nodes comprise of the labeled and unlabeled examples of the dataset.

c. The edges are generally weighted and undirected and it is assumed that the examples connected by heavy edges have the same label.

d. The edge weight w_{ij} reflects how close the two nodes x_i and x_j are. The heavier the edge, the closer they are to each other.

5. **Semi-Supervised Support Vector Machines (S3VMs) :**

a. Semi-supervised support vector machines can be thought of as an extension of support vector machines with unlabeled data.

b. In a standard support vector machine, labeled data is used to find a maximum margin linear boundary in the reproducing kernel Hilbert Space.

c. In an S3VM, the unlabeled data guides the placement of the decision boundary.

d. Labeled data is used to find a labeling of the unlabeled data, so that a linear boundary has the maximum distance from both the original labeled data and the unlabeled data.

e. The assumption in this model is that the decision boundary is situated in a low density region, between two classes $y \in [-1, 1]$.

f. S3VMs can be viewed as SVM with an additional regularization term for the unlabeled data.

Que 2.20. **Write the advantages and disadvantages of following semi-supervised learning algorithm :**

1. **Self training**
2. **Generative models**
3. **Co-training**
4. **Graph based algorithms**
5. **Semi-supervised support vector machines (S3VMs).**

Answer

1. **Self training :**

Advantages :

1. Simplest of all semi-supervised learning algorithms.
2. It is a wrapper method and applies to almost all existing classifiers.

Disadvantages :

1. Mistakes reinforce or strengthen themselves.
2. In terms of convergence, cannot give too much information.

2. Generative models :

Advantages :

1. If the model is close to correct, it can give efficient predictions.
2. The knowledge of the structure of the problem or data can be included by modelling it.

Disadvantages :

1. They often do not provide good solutions to classification problems.
2. There can be a problem balancing the impact of labeled and unlabeled data when the unlabeled data is much more than labeled data.
3. Local optima of the EM algorithm.
4. Modelling effort is much more demanding than discriminative models.
5. Since generative models are very precise, there is a high likelihood of them being incorrect.
6. Unlabeled data will hurt the prediction if the model is wrong.

3. Co-training :

Advantages :

1. It is a wrapper method. It can use any classifier.
2. Less susceptible to mistakes than self training.

Disadvantages :

1. The feature set might not be able to split.

4. Graph based algorithms :

Advantages :

1. Lucid mathematical framework.
2. Good performance if the graph fits the task.
3. It can be applied in directed graphs.

Disadvantages :

1. Bad performance if graph does not fit the task.
2. Performance is vulnerable to graph structure and edge weights.

5. Semi-Supervised Support Vector Machines (S3VMs) :

Advantages :

1. They are valid wherever support vector machines are valid.
2. Lucid mathematical framework.

Disadvantages :

1. Optimization is difficult since algorithm can be caught in bad local optima.

☺☺☺

3 UNIT

Dimensionality Reduction

CONTENTS

PART-1

Linear (PCA, LDA) and Manifolds.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.1. Write a short note on Principal Component Analysis (PCA).

Answer

1. Principal Component Analysis (PCA) technique adopted for dimensionality reduction.
2. Using the PCA technique, a higher dimensional data space can be transformed onto a lower dimensional space. This transformation is also called the Hotelling transform.
3. It linearly transforms a high-dimensional input vector into a low-dimensional one whose components are uncorrelated through the calculation of eigen vectors of the covariance matrix of the original inputs.
4. The primary advantages of the PCA are the reduction of the dimensionality of the data set and the identification of new meaningful underlying variables.
5. The main issue of a principal component analysis is to reveal the true dimensionality of the space in which the data lie.
6. The goal of principal component analysis is to identify the most meaningful basis to re-express a data set.
7. In other words, the PCA technique consists of finding uncorrelated linear transformations, $y_1, y_2, y_3, \ldots, y_p$ of the original components, $x_1, x_2, x_3, \ldots, x_p$ such that the y components are chosen in such a way that y_1 has maximum variance ; y_2 has maximum variance subject to being uncorrelated with y_1, and so forth.
8. The very first step of the PCA algorithm will be to normalize the components so that they have zero mean and unity variance.
9. Then, an orthogonalization method can be used to compute the principal components of the normalized components.
10. **PCA algorithm :**

Step 1 : Get data.

Step 2 : Subtract the mean.

Step 3 : Calculate the covariance matrix.

Step 4 : Calculate the eigen vectors and eigen values of the covariance matrix.

Step 5 : Choosing components and forming a feature vector.

Step 6 : Deriving the new data set, this is the final step in PCA, and is also the easiest.

Once we have chosen the components (eigen vectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

11. PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, known as principal space, such that the variance of the projected data is maximized.
12. Equivalently, it can be defined as the mean squared distance between the data points and their projections. The process of orthogonal projection is illustrated in the Fig. 3.1.1

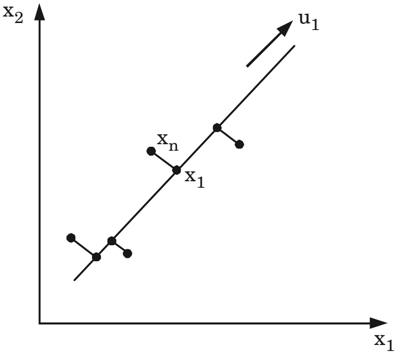


Fig. 3.1.1. PCA as an orthogonal projection.

Que 3.2. **What are the advantages and disadvantages of PCA ?**

Answer

Advantages of PCA are :

1. Lack of redundancy of data given the orthogonal components.
2. Reduced complexity in images grouping with the use of PCA.
3. Smaller database representation since only the trainee images are stored in the form of their projections on a reduced basis.
4. Reduction of noise since the maximum variation.

Disadvantages of PCA are :

1. The covariance matrix is difficult to be evaluated in an accurate manner.

2. Even the simplest invariance could not be captured by the PCA unless the training data explicitly provides this information.

Que 3.3. **What are the features of PCA ?**

Answer

S. No.	Feature	Principal component analysis
1.	Discrimination between classes	PCA manages the entire data for the principal components analysis without taking into consideration the fundamental class structure.
2.	Applications	PCA applications in the significant fields of criminal investigation are beneficial.
3.	Computation for large datasets	PCA does not require large computations.
4.	Direction of maximum discrimination	The directions of the maximum discrimination are not the same as the directions of maximum variance as it is not required to utilize the class information such as the within class scatter and between class scatter.
5.	Focus	PCA examines the directions that have widest variations.
6.	Supervised learning technique	PCA is an unsupervised technique.
7.	Well distributed classes in small datasets	PCA is not as powerful as other methods.

Que 3.4. **Write a short note on Linear Discriminant Analysis (LDA).**

Answer

1. Linear Discriminant Analysis (LDA) is a technique used for data classification and dimensionality reduction.
2. Linear discriminant analysis easily handles the case where the values within class frequencies are unequal and their performances have been examined on randomly generated test data.

3. This method maximizes the ratio of between class variance to within the class variance in any particular data set, thereby guaranteeing maximal separability.
4. The use of linear discriminant analysis for data classification is applied to a classification problem in speech recognition.
5. LDA works when the measurement made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminate correspondence analysis.
6. LDA is closely related to ANOVA (Analysis of Variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements.
7. LDA is also closely related to Principal Component Analysis (PCA) and factor analysis for linear combinations of variables which best explains the data.

Difference between PCA and LDA :

1. The prime difference between PCA and LDA is that PCA does feature classification and LDA does data classification.
2. In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA does not change the location but only tries to provide more class separability and draw a decision region between the given classes. This method also helps to better understand the distribution of the feature data.

Que 3.5. **Explain LDA for two classes.**

Answer

1. Consider a set of observations $\vec{x}$ (also called features, attributes, variables or measurements) for each sample of an object or event with known class y. This set of samples is called the training set.
2. The classification problem is then to find a good predictor for the class y of any sample of the same distribution (not necessarily from the training set) given only an observation $\vec{x}$.
3. LDA approaches the problem by assuming that the conditional probability density functions $p(\vec{x}\ |y = 0)$ and $p(\vec{x}\ |\ y = 1)$ are both normally distributed with mean and covariance parameters $\left(\vec{\mu}_0 \sum_{Y=0}\right)$ and $\left(\vec{\mu}_1 \sum_{Y=1}\right)$, respectively.

4. Under this assumption, the Bayes optimal solution is to predict points as being from the second class if the ratio of the log-likelihoods is below some threshold T, so that :

$$(\vec{x}-\mu_0)^T \sum_{Y=0}^{-1} (\vec{x}-\vec{\mu}_0) + 1n\left|\sum_{y=0}\right| - (\vec{x}-\vec{\mu}_1)^T \sum_{Y=1}^{-1} (\vec{x}-\vec{\mu}_1) - 1n\left|\sum_{Y=1}\right| < T$$

Without any further assumptions, the resulting classifier is referred to as QDA (quadratic discriminate analysis).

5. LDA also makes the simplifying homoscedastic assumption (*i.e.*, that the class covariance's are identical, so $\sum_{Y=0} = \sum_{Y=1} = \Sigma$) and that the covariance's have full rank.

6. In this case, several terms cancel and the above decision criterion becomes a threshold on the dot product.

$$\vec{w} \cdot \vec{x} < c$$

For some threshold constant c, where

$$w = \sum^{-1}\left(\vec{\mu}_1 - \vec{\mu}_0\right)$$

7. This means that the criterion of an input $\vec{x}$ being in a class y is purely a function of this linear combination of the known observations.

8. It is often useful to see this conclusion in geometrical terms : the criterion of an input $\vec{x}$ being in a class y is purely a function of projection of multidimensional-space point $\vec{x}$ onto direction $\vec{w}$.

9. In other words, the observation belongs to y if corresponding $\vec{x}$ is located on a certain side of a hyperplane perpendicular to $\vec{w}$. The location of the plane is defined by the threshold c.

Que 3.6. **What are the advantages and disadvantages of LDA ?**

Answer

Advantages :

1. **Completely unsupervised :** It can learn topics without the need for annotated training data.
2. **Intuitive :** Forum threads can intuitively be thought of a document in a corpus.
3. **Built-in classification :** Documents are distributions over topics. It can classify documents by high probability topics.

Disadvantages :

1. **Not scalable :** Each global topic update requires one full pass over the corpus. Entire corpus must fit in main memory. Not feasible for us.
2. **Inefficient to update a model :** New threads are constantly being created. Want to update model, not re-run it on entire corpus.

Que 3.7. **What are the applications of LDA ?**

Answer

Following are the application LDA :

1. **Face recognition :**
   a. In the field of Computer Vision, face recognition is a very popular application in which each face is represented by a very large number of pixel values.
   b. Linear Discriminant Analysis (LDA) is used here to reduce the number of features to a more manageable number before the process of classification.
   c. Each of the new dimensions generated is a linear combination of pixel values, which form a template. The linear combinations obtained using Fisher's linear discriminant are called Fisher faces.
2. **Medical :**
   a. In this field, Linear Discriminant Analysis (LDA) is used to classify the patient disease state as mild, moderate or severe based upon the patient various parameters and the medical treatment he is going through.
   b. This helps the doctors to reduce the pace of their treatment.
3. **Customer identification :**
   a. Suppose we want to identify the type of customers which are most likely to buy a particular product in a shopping mall.
   b. By doing a simple question and answers survey, we can gather all the features of the customers.
   c. Here, linear discriminant analysis will help us to identify and select the features which can describe the characteristics of the group of customers that are most likely to buy that particular product in the shopping mall.

Que 3.8. **Explain manifold learning.**

Answer

1. A manifold is a connected region. Mathematically, it is a set of points, associated with a neighborhood around each point.

2. From any given point, the manifold locally appears to be a Euclidean space.
3. The definition of a neighborhood surrounding each point implies the existence of transformations that can be applied to move on the manifold from one position to a neighboring one.
4. In the example of the world surface as a manifold, one can walk north, south, east, or west.
5. Although there is a formal mathematical meaning to the term "manifold," in machine learning it tends to be used more loosely to designate a connected set of points that can be approximated well by considering only a small number of degrees of freedom, or dimensions, embedded in a higher-dimensional space.
6. Each dimension corresponds to a local direction of variation.
7. Fig. 3.8.1 shows an example of training data lying near a one-dimensional manifold embedded in two-dimensional space.
8. In the context of machine learning, we allow the dimensionality of the manifold to vary from one point to another.
9. This happens when a manifold intersects itself. For example, a Fig. 3.8.1 is a manifold that has a single dimension in most places but two dimension at the intersection at the center.

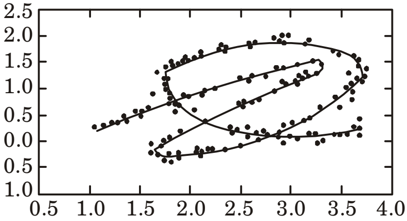


Fig. 3.8.1. Data sampled from a distribution in a two-dimensional space that is actually concentrated near a one-dimensional manifold, like a twisted string. The solid line indicate the underlying manifold that the learner should infer.

10. Many machine learning problems seem hopeless if we expect the machine learning algorithm to learn functions with interesting variations across all of R^n.
11. Manifold learning algorithms surmount this obstacle by assuming that most of R^n consists of invalid inputs, and that interesting inputs occur only along a collection of manifolds containing a small subset of points, with interesting variations in the output of the learned function occurring only along directions that lie on the manifold, or with interesting variations happening only when we move from one manifold to another.

12. Manifold learning was introduced in the case of continuous valued data and the unsupervised learning setting, although this probability concentration idea can be generalized to both discrete data and the supervised learning setting, the key assumption remains that probability mass is highly concentrated.
13. The assumption that the data lies along a low-dimensional manifold may not always be correct or useful.

PART-2

Metric Learning, Autoencoders and Dimensionality Reduction in Networks.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.9. **What do you understand by metric learning ?**

Answer

1. Metric learning aims at automatically constructing task-specific distance metrics from (weakly) supervised data, in a machine learning manner.
2. The learned distance metric can then be used to perform various tasks (for example, *k*-NN classification, clustering, information retrieval).
3. Metric learning problems fall into two main categories depending on the type of supervision available about the training data :
   a. **Supervised learning :**
      i. The algorithm has access to a set of data points, each of them belonging to a class (label) as in a standard classification problem.
      ii. The goal is to learn a distance metric that puts points with the same label close together while pushing away points with different labels.
   b. **Weakly supervised learning :**
      i. The algorithm has access to a set of data points with supervision only at the tuple level (typically pairs, triplets, or quadruplets of data points).
      ii. A classic example of such weaker supervision is a set of positive and negative pairs: in this case, the goal is to learn a distance metric that puts positive pairs close together and negative pairs far away.

4. Based on weakly supervised data, the metric learning problem is generally formulated as an optimization problem where one seeks to find the parameters of a distance function that optimize some objective function measuring the agreement with the training data.

Que 3.10. **Write short note on nearest neighbour rule.**

Answer

1. In pattern recognition, the *k*-nearest neighbours algorithm (*k*-NN) is a method for classifying objects based on closest training examples in the feature space.
2. *k*-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.
3. The *k*-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms in which an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its *k* nearest neighbours (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbour.
4. Nearest neighbour rules in effect compute the decision boundary in an implicit manner.
5. The available variables are divided into the explanatory variables (x) and the target variable (y).
6. A sample of observations in the form (x, y) is collected to form a training data set.
7. For this training data, a distance function is introduced between the x values of the observations.
8. This can be used to define, for each observation, a neighbourhood formed by the observations that are closest to it, in terms of the distance between the x values.
9. For a continuous response variable, the nearest-neighbour fitted value for each observation's response value y_i is defined by :

$$y_i = \frac{1}{k} \sum_{x_j \in N(x_i)} y_j$$

10. Nearest-neighbour methods can also be used for predictive classification.

Nearest-Neighbour algorithm :

```
1. begin
2. initialize $c$; $c' = n$; $D_i = \{x_i\}$; $i = 1, ......, n$
3. do
4. $c' = c' - 1$
```

5. Find nearest clusters D_i and D_j
6. Merge D_i and D_j
7. until $c = c'$
8. return c clusters
9. end

a. To find the nearest clusters in step 5, the following clustering criterion function is used :

   $d_{\min}(D_i, D_j) = \min ||x - x'||$, where $x \varepsilon D_i$ and $x' \varepsilon D_j$

b. The merging of the two clusters in step 6 simply corresponds to adding an edge between the nearest pair of nodes in D_i and D_j.

Que 3.11. **What is clustering ? Describe *k*-mean clustering technique.**

Answer

Clustering :

1. Clustering is the process or grouping of classifying objects on the basis of a close association or shared characteristics.
2. The objects can be physical or abstract entities, and the characteristics can be attribute values, relations among the objects, and combinations of both.
3. At a more abstract level, the objects might be some concept such as the quality of the items purchased. The classification in this case might be made on the basis of some subjective criteria, such as poor, average, or good.
4. Clustering is essentially a discovery learning process in which similarity patterns are found among a group of objects.

***k*-mean clustering :**

1. The basic idea of k-means clustering is that clusters of items with the same target category are identified and predictions for new data items are made by assuming that they are of the same type as the nearest cluster center.
2. Suppose we have n feature vectors $X_1, X_2, \ldots\ldots, X_n$, all belonging to the same class C and we know that they belong to k clusters such that $k < n$.
3. If clusters are well separated we can use a minimum distance classifier to separate them.
4. We first initialize the means $\mu_1, \ldots\ldots, \mu_k$ of k clusters. One of the ways to do this is just to assign random numbers to them.
5. We then determine the membership of each X by taking the $||X - \mu_i||$.
6. The minimum distance determines X's membership in a respective cluster. This is done for all n feature vectors.

Algorithm : The algorithm is composed of the following steps :

1. Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the k centroids.
4. Repeat step 2 and step 3 until the centroids no longer move.
   a. Fig. 3.11.1 (*a*) explains how the algorithm works in the case when $k = 2$. We randomly select the values for μ_1 and μ_2, and the algorithm converges when the means no longer change. (Let $\mu_i = m_i$)
   b. There are certain problems with this algorithm, like when the means are chosen such that some M_i happens to be 0, so that it cannot be updated.
   c. It is, hence, advisable to try a number of different starting points. The results depend a lot on the value of k, *i.e.*, the actual number of clusters. Often, we have no way of knowing how many cluster exist.

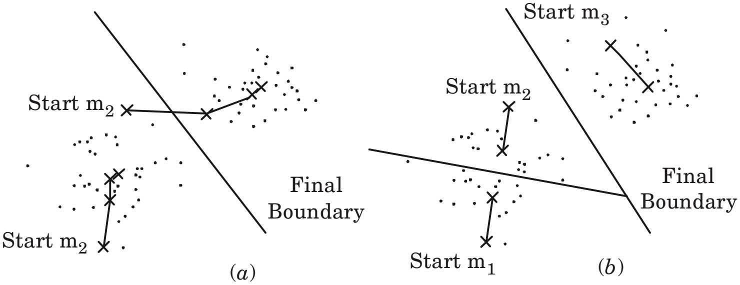


Fig. 3.11.1.

   d. Fig. 3.11.1(*b*) shows what happens when we use $k = 3$. (Let $\mu_i = m_i$). Sometimes the clustering division turns out to be better at higher k.
   e. We can go all the way upto $k = n$, this procedure will give us something known as the nearest neighbour classifier.
   f. It performs great if the number of feature vector is large; however, computationally it is much more expensive.

Que 3.12. Discuss autoencoders.

Answer

1. An autoencoder is a neural network that is trained to attempt to copy its input to its output.

2. Internally, it has a hidden layer h that describes a code used to represent the input.
3. The network may be viewed as consisting of two parts *i.e.*, an encoder function $h = f(x)$ and a decoder that produces a reconstruction $r = g(h)$.
4. If an autoencoder succeeds in simply learning to set $g(f(x)) = x$ everywhere, then it is not especially useful.
5. Autoencoders are designed to be unable to learn to copy perfectly.
6. Usually they are restricted to copy only input that resembles the training data.
7. Because the model is forced to prioritize which aspects of the input should he copied, it often learns useful properties of the data.
8. Modern autoencoders have generalized the idea of an encoder and a decoder beyond deterministic functions to stochastic mappings p_{encoder} $(h \mid x)$ and p_{decoder} $(x \mid h)$.
9. Autoencoders were used for dimensionality reduction or feature learning.
10. Unlike general feedforward networks, autoencoders may also be trained using recirculational learning algorithm based on comparing the activations of the network on the original input to the activations on the reconstructed input.
11. Recirculation is regarded as more biologically plausible than back-propagation, but is rarely used for machine learning applications.

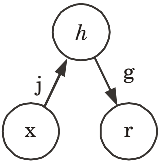


Fig. 3.12.1. The general structure of an autoencoder.

Que 3.13. **Explain dimensionality reduction in networks.**

Answer

1. Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration.
2. Dimensionality reduction methods include wavelet transforms and principal components analysis, which transform or project the original data onto a smaller space.
3. Attribute subset selection is a method of dimensionality reduction in which irrelevant, weakly relevant or redundant attributes or dimensions are detected and removed.
4. The goal of dimensionality reduction methods is to remove redundant objectives such that its main features are preserved to the extent possible.

5. Dimensionality reduction may be both linear or non-linear, depending upon the method used.

Que 3.14. **What are the components of dimensionality reduction ?**

Answer

There are two components of dimensionality reduction :

1. **Feature selection :** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways :
   a. Filter
   b. Wrapper
   c. Embedded
2. **Feature extraction :** This reduces the data in a high dimensional space to a lower dimension space, *i.e.*, a space with lesser number of dimensions.

Que 3.15. **What are the various methods of dimensionality reduction ?**

Answer

The various methods used for dimensionality reduction include :

a. **Principal Component Analysis (PCA) :** Refer Q. 3.1, Page 3–2M, Unit-3.

b. **Linear Discriminant Analysis (LDA) :** Refer Q. 3.4, Page 3–4M, Unit-3.

c. **Generalized Discriminant Analysis (GDA) :**
   1. Linear discriminant analysis (LDA) is a traditional statistical method which has proven successful on classification problems.
   2. The procedure is based on an eigenvalue resolution and gives an exact solution of the maximum of the inertia. But this method fails for a nonlinear problem.
   3. To overcome this limitation Generalized Discriminant Analysis (GDA) is developed by mapping the input space into a high dimensional feature space with linear properties.
   4. In the new space, one can solve the problem in a classical way such as the LDA method.
   5. The main idea is to map the input space into a convenient feature space in which variables are nonlinearly related to the input space.
   6. Generalized Discriminant Analysis is use to deal with nonlinear discriminant analysis using kernel function operator.
   7. Kernel discriminants are greatly used because :

   i. They permit to establish nonlinear boundaries between classes and
   ii. They offer the possibility of visualizing graphically the data vectors belonging to different classes.

8. GDA operates on a kernel matrix of size N × N, (N denotes the sample size) and is for large N prohibitive.

Que 3.16. **What are the advantages and disadvantages of dimensionality reduction ?**

Answer

Advantages of dimensionality reduction :

1. It helps in data compression, and hence reduced storage space.
2. It reduces computation time.
3. It also helps to remove redundant features.

Disadvantages of dimensionality reduction :

1. It may lead to some amount of data loss.
2. It finds linear correlations between variables, which is sometimes undesirable.
3. It fails in cases where mean and covariance are not enough to define datasets.
4. We may not know how many principal components to keep- in practice, some thumb rules are applied.

PART-3

Introduction to ConvNet, Architectures :
AlexNet, VGG, Inception, ResNet.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.17. **Write short note on ConvNet.**

Answer

Refer Q. 2.14, Page 2–11M, Unit-2.

Que 3.18. **Explain AlexNet architecture.**

Answer

1. AlexNet solves the problem of image classification where the input is an image of one of 1000 different classes (for example, cats, dogs etc.) and the output is a vector of 1000 numbers.
2. The i^{th} element of the output vector is interpreted as the probability that the input image belongs to the i^{th} class. Therefore, the sum of all elements of the output vector is 1.
3. The input to AlexNet is an RGB image of size 256×256. This means all images in the training set and all test images need to be of size 256×256.
4. AlexNet was larger than previous CNNs used for computer vision tasks. It has 60 million parameters and 650,000 neurons and took five to six days to train on two GTX 580 3GB GPUs.

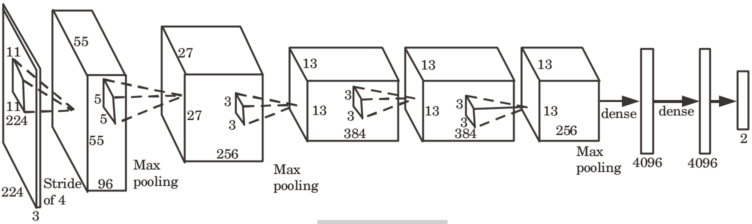


Fig. 3.18.1.

5. This architecture was one of the first deep networks to push ImageNet classification accuracy by a significant stride in comparison to traditional methodologies.
6. It is composed of five convolutional layers followed by three fully connected layers, as depicted in Fig 3.18.1.
7. AlexNet uses ReLu (Rectified Linear Unit) for the non-linear part, instead of a Tanh or Sigmoid function which was the earlier standard for traditional neural networks. ReLu is given by

$$f(x) = \max(0, x)$$

8. The advantage of the ReLu over sigmoid is that it trains much faster because the derivative of sigmoid becomes very small in the saturating region and therefore the updates to the weights almost vanish (Figure 3.18.2). This is called vanishing gradient problem.
9. In the network, ReLu layer is put after each and every convolutional and Fully-Connected layers (FC).

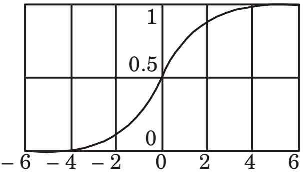


Fig. 3.18.2.

Que 3.19. **What is VGG ?**

Answer

1. This architecture makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3X3 kernel-sized filters one after another.
2. With a given receptive field (the effective area size of input image on which output depends), multiple stacked smaller size kernel is better than the one with a larger size kernel because multiple non-linear layers increases the depth of the network which enables it to learn more complex features, and that too at a lower cost.
3. For example, three 3X3 filters on top of each other with stride 1 ha a receptive size of 7, but the number of parameters involved is 3*(9C^2) in comparison to 49C^2 parameters of kernels with a size of 7.
4. Here, it is assumed that the number of input and output channel of layers is C. Also, 3X3 kernels help in retaining finer level properties of the image.
5. We can see that in VGG-D, there are blocks with same filter size applied multiple times to extract more complex and representative features.
6. This concept of blocks/modules became a common theme in the networks after VGG.
7. The VGG convolutional layers are followed by three fully connected layers. The width of the network starts at a small value of 64 and increases by a factor of two after every sub-sampling/pooling layer.

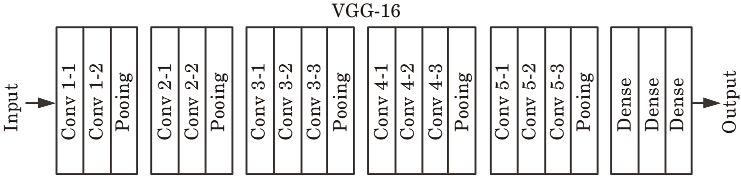


Fig. 3.19.1.

Que 3.20. **Discuss the Inception architecture.**

Answer

1. The GoogLeNet builds on the idea that most of the activations in a deep network are either unnecessary (value of zero) or redundant because of correlations between them.
2. Therefore, the most efficient architecture of a deep network will have a sparse connection between the activations, which implies that all 512

output channels will not have a connection with all the 512 input channels.

3. There are techniques to prune out such connections which would result in a sparse weight/connection.
4. But kernels for sparse matrix multiplication are not optimized in BLAS or CuBlas (CUDA for GPU) packages which render them to be slower than their dense counterparts.
5. So GoogLeNet devised a module called inception module that approximates a sparse CNN with a normal dense construction.
6. Since only a small number of neurons are effective as mentioned earlier, the width/number of the convolutional filters of a particular kernel size is kept small. Also, it uses convolutions of different sizes to capture details at varied scales (5X5, 3X3, 1X1).

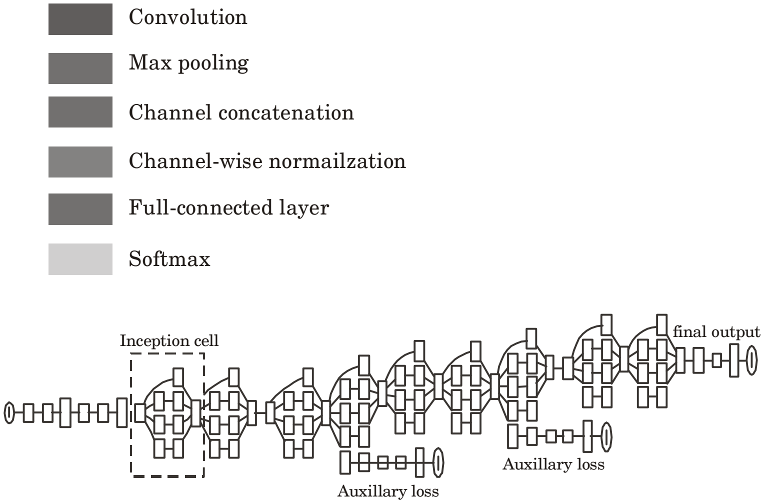


Fig. 3.20.1.

Que 3.21. **Explain ResNet architecture in detail.**

Answer

1. At the ILSVRC 2015, the so-called Residual Neural Network (ResNet) by Kaiming He et al introduced a novel architecture with "skip connections" and features heavy batch normalization.
2. Such skip connections are also known as gated units or gated recurrent units and have a strong similarity to recent successful elements applied in RNNs.

3. They were able to train a NN with 152 layers while still having lower complexity than VGGNet.
4. It achieves a top-5 error rate of 3.57% which beats human-level performance on this dataset.

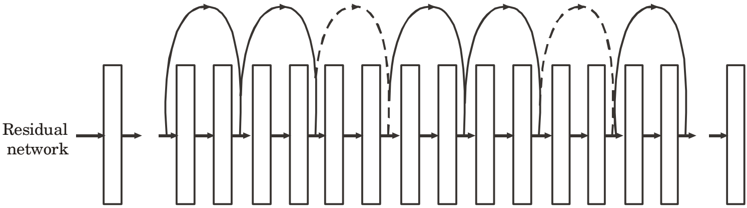


Fig. 3.21.1.

PART-4

Training a ConvNet : Weights initialization, Batch Normalization, Hyperparameter Optimization.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.22. **Explain training of convolutional architecture.**

Answer

1. During network training, the filter weights are adjusted, so as to improve the classification performance of the network.
2. This can be done using a method called backpropagation, where the gradient of an error function is computed with respect to all network weights, going all the way to the input connections of the network.
3. Network weights are updated by the following equation relating the step to the gradient and the learning rate, denoted η.

$$W_{new} = W - \eta \frac{dE}{dW} \qquad ...(3.22.1)$$

4. An error function can be expressed as a sum of squared differences between the network's output and the correct output, over all discrete points in the output.

5. This sort of scoring function works for cases where the network output is a vector, matrix, or tensor of continuous real values.

$$E(W, b) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}||hW,b(I^{(i)}) - y^{(i)}||^2 \qquad ...(3.22.2)$$

6. The gradient of this scoring function would be taken in eq. (3.22.1)
7. For classification network, the scoring function is computed differently.
8. Because the output is a one-of-N vector, where the highest value component represents the network's best estimation of the image class (*i.e.*, car, boat, airplane, etc), an error function for categorical output is needed.
9. This is called the categorical cross entropy.

Que 3.23. **Explain weight initialization.**

Answer

1. When we create our neural networks, we have to make choices for the initial weights and biases.
2. We have been choosing them according to a prescription that prescription was to choose both the weights and biases using independent Gaussian random variables, normalized to have mean 0 and standard deviation 1.
3. While this approach has worked well, it was quite ad hoc, and it's worth revisiting to see if we can find a better way of setting our initial weights and biases, and perhaps help our neural networks learn faster.
4. It turns out that we can do quite a bit better than initializing with normalized Gaussians.
5. For example, suppose we are working with a network with a large number say 1,000 of input neurons.

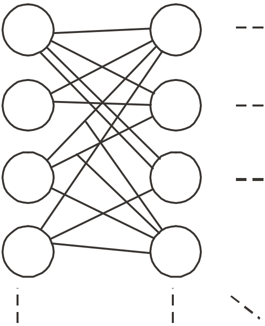


Fig. 3.23.1.

6. And let's suppose we have used normalized Gaussians to initialize the weights connecting to the first hidden layer.
7. To concentrate specifically on the weights connecting the input neurons to the first neuron in the hidden layer and ignore the rest of the network.
8. We will suppose that we are trying to train using a training input x in which half the input neurons are on, *i.e.*, set to 1, and half the input neurons are off, *i.e.*, set to 0.
9. Let's consider the weighted sum $z = \Sigma_f w_j x_j + b$ of inputs to our hidden neuron, 500 terms in this sum vanish, because the corresponding input x_f is zero.
10. And so z is a sum over a total of 501 normalized Gaussian random variables, accounting for the 500 weight terms and the one extra bias term.
11. Thus z is itself distributed as a Gaussian with mean zero and standard deviation $\sqrt{501} \approx 22.4$.
12. That is, z has a very broad Gaussian distribution, not sharply peaked at all :

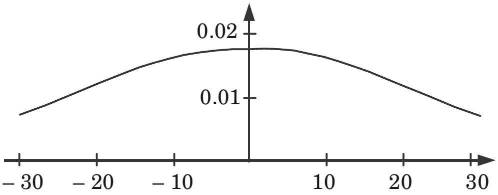


Fig. 3.23.2.

13. In particular, we can see from this graph that it's quite likely that $|z|$ will be pretty large, *i.e.*, either $z > 1$ or $z - 1$.
14. If that's the case then the output $\sigma\{z\}$ from the hidden neuron will be very close to either 1 or 0.
15. That means our hidden neuron will have saturated.
16. And when that happens, as we know, making small changes in the weights will make only absolutely miniscule changes in the activation of our hidden neuron.
17. That miniscule change in the activation of the hidden neuron will, in turn, barely affect the rest of the neurons in the network at all, and we will see a correspondingly miniscule change in the cost function.
18. As a result those weights will only learn very slowly when we use the gradient descent algorithm.

Que 3.24. **What is batch normalization ?**

Answer

Refer Q. 2.10, Page 2–8M, Unit-2.

Que 3.25. **Explain hyperparameter optimization.**

Answer

1. Most machine learning algorithms have several settings that we can use to control the behavior of the learning algorithm.
2. These settings are called hyperparameters.
3. The values of hyperparameters are not adapted by the learning algorithm itself (though we can design a nested learning procedure where one learning algorithm learns the best hyperparameters for another learning algorithm).
4. In polynomial regression, there is a single hyperparameter *i.e.*, the degree of the polynomial, which acts as a capacity hyperparameter.
5. The λ value used to control the strength of weight decay is another example of a hyperparameter.
6. Sometimes a setting is chosen to be a hyperparameter that the learning algorithm does not learn because it is difficult to optimize.
7. More frequently, we do not learn the hyperparameter because it is not appropriate to learn that hyperparameter on the training set.
8. This applies to all hyperparameters that control model capacity.
9. If learned on the training set, such hyperparameters would always choose the maximum possible model capacity, resulting in overfitting.
10. For example, we can always fit the training set better with a higher degree polynomial and a weight decay setting of $\lambda = 0$ than we could with a lower degree polynomial and a positive weight decay setting.
11. To solve this problem, we need a validation set of examples that the training algorithm does not observe.
12. It is important that the test examples are not used in any way to make choices about the model, including its hyperparameters.
13. For this reason, no example from the test set can be used in the validation set.
14. Therefore, we always construct the validation set from the training data.
15. Specifically, we split the training data into two disjoint subsets.
16. One of these subsets is used to learn the parameters.
17. The other subset is our validation set, used to estimate the generalization error during or after training, allowing for the hyperparameters to be updated accordingly.

18. The subset of data used to learn the parameters is still typically called the training set, even though this may be confused with the larger pool of data used for the entire training process.
19. The subset of data used to guide the selection of hyperparameters is called the validation set.
20. Typically, one uses about 80 % of the training data for training and 20 % for validation.
21. Since the validation set is used to train the hyperparameters, the validation set error will underestimate the generalization error, though typically by a smaller amount than the training error.
22. After all hyperparameter optimization is complete, the generalization error may be estimated using the test set.

☺☺☺

4

UNIT

Optimization and Generalization

CONTENTS

PART-1

Optimization in Deep Learning, Non-Convex Optimization for Deep Networks, Stochastic Optimization Generalization in Neural Networks.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.1. **What do you mean by optimization in deep learning ?**

Answer

1. Optimization refers to the task of minimizing some function $f(x)$ by altering x.
2. We usually phrase most optimization problems in terms of minimizing $f(x)$. Maximization may be accomplished via a minimization algorithm by minimizing $f(x)$.
3. The function we want to minimize or maximize is called the objective function or criterion.
4. When we are minimizing it, we may also call it the cost function or loss function.
5. We often denote the value that minimizes or maximizes a function with a superscript. For example, we might say $x = \arg\min f(x)$.
6. Optimization algorithms that use only the gradient, such as gradient descent, are called first-order optimization algorithms.
7. Optimization algorithms that use the Hessian matrix, such as Newton's method, are called second-order optimization algorithms.

Que 4.2. **Describe optimization method used in deep learning.**

Answer

1. Optimization methods can be classified into general methods and methods tailored for a specific class of problems.
2. Specific methods such as linear programming and quadratic programming are more efficient than the general methods in solving the problems because they are tailored for it.

3. However, they are not applicable to general problems.
4. General methods can be divided to local optimization methods and global optimization methods.
5. Local optimization methods only provide results that are locally optimal. However, their computational cost is lower than those of global search methods.
6. Newton method and sequential quadratic programming are examples of local optimization methods.
7. Global optimization methods are heuristic-based methods. This means that there is no guarantee for their result to be globally optimal.
8. Genetic algorithm (GA) and simulated annealing are the examples of methods that do not have any restriction in the type of functions that are used in stating the objective and constraint functions.

Que 4.3. **Explain optimization algorithm.**

Answer

1. Optimization algorithms helps us to minimize (or maximize) an objective function (loss function) $E(x)$ which is simply a mathematical function dependent on the model's internal learnable parameters which are used in computing the target values (Y) from the set of predictors (X) used in the model.
2. For example, we call the Weights (W) and the Bias (B) values of the neural network as its internal learnable parameters which are used in computing the output values and are learned and updated in the direction of optimal solution *i.e.*, minimizing the loss by the network's training process and also play a major role in the training process of the Neural Network model.
3. There are two types of optimization algorithm :

   a. **First-order optimization algorithms :**

      i. These algorithms minimize or maximize a loss function $E(x)$ using its Gradient values with respect to the parameters.

      ii. Most widely used first-order optimization algorithm is Gradient descent. The first-order derivative tells us whether the function is decreasing or increasing at a particular point.

      iii. First-order derivative give us a line which is tangential to a point on its error surface.

   b. **Second order optimization algorithms :**

      i. Second-order methods use the second-order derivative which is also called Hessian to minimize or maximize the loss function.

ii. The Hessian is a matrix of second-order partial derivatives. Since the second derivative is costly to compute, the second order is not used much.

iii. The second order derivative tells us whether the first derivative is increasing or decreasing which hints at the function's curvature.

iv. Second-order derivative provide us with a quadratic surface which touches the curvature of the error surface.

Que 4.4. **What is stochastic optimization ?**

Answer

1. Stochastic optimization refers to the collection of methods for minimizing or maximizing an objective function when randomness is present.
2. Randomness usually enters the problem in two ways *i.e.*, through the cost function or the constraint set.
3. Optimization refers to any optimization method that employs randomness within communities, we only consider those settings where the objective function or constraints are random.
4. The most prominent division is between solution methods for problems with a single time period (single stage problems) and those with multiple time periods (multistage problems).
5. Single stage problems try to find a single optimal decision, such as the best set of parameters for a statistical model given data.
6. Multistage problems try to find an optimal sequence of decisions, such as scheduling water releases from hydroelectric plants over a two year period.
7. Single stage problems are solved with modified deterministic optimization methods.
8. However, the dependence of future decisions on random outcomes makes direct modification of deterministic methods difficult in multistage problems.
9. Multistage methods are more reliant on statistical approximation and strong assumptions about problem structure, such as finite decision and outcome spaces, or a compact Markovian representation of the decision process.

Que 4.5. **Discuss generalization in neural network.**

Answer

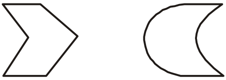
1. Generalization of the ANN is ability to handle unseen data.

2. The generalization capability of the network is mostly determined by system complexity and training of the network.
3. Poor generalization is observed when the network is over-trained or system complexity (or degree of freedom) is relatively more than the training data.
4. A smaller network which can fit the data will have the *k* good generalization ability.
5. Network parameter pruning is the methods used to reduce the degree of freedom of a network and hence improve its generalization.
6. It is important to estimate the improvement in generalization and rate of improvement as pruning being incorporated in the network.
7. A method is developed in this research to evaluate generalization capability and rate of convergence towards the generalization.
8. Using the proposed method, experiments have been conducted to evaluate Multi-Layer Perceptron neural network with pruning being incorporated for handwritten numeral recognition.

Que 4.6. **Explain non-convex optimization in detail.**

Answer

1. Non-convex optimization involves a function which has multiple optima, from which only one is global optima.
2. Depending on the loss surface, it can be very difficult to locate the global optima.
3. A non-convex optimization problem is any problem where the objective or any of the constraints are non-convex, as shown in Fig. 4.6.1.



Non-convex

Fig. 4.6.1.

4. Such a problem may have multiple feasible regions and multiple locally optimal points within each region.
5. It can take time exponential in the number of variables and constraints to determine that a non-convex problem is infeasible, that the objective function is unbounded, or that an optimal solution is the “global optimum” across all feasible regions.

PART-2

Spatial Transformer Networks, Recurrent Networks, LSTM, Recurrent Neural Network Language Models.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.7. **What do you understand by spatial transformer networks ?**

Answer

1. Spatial transformer networks are the generalization of differentiable attention to any spatial transformation.
2. Spatial transformer networks allow a neural network to learn how to perform spatial transformations on the input image in order to enhance the geometric invariance of the model.
3. For example, it can crop a region of interest, scale and correct the orientation of an image.
4. It can be a useful mechanism because CNNs are not invariant to rotation, scale and general affine transformations.
5. One of the best things about STN (Spatial Transform Network) is the ability to simply plug it into any existing CNN with very small modification.
6. Following are the components of spatial transform network :
   a. **Localization network :** The localization network is a regular CNN which regress the transformation parameters. The transformation is never learned explicitly from this dataset, instead the network learns automatically the spatial transformations that enhances the global accuracy.
   b. **Grid generator :** The grid generator generates a grid of coordinates in the input images corresponding to each pixel from the output image.
   c. **Sampler :** The sampler uses the parameters of the transformation and applies it to the input image.

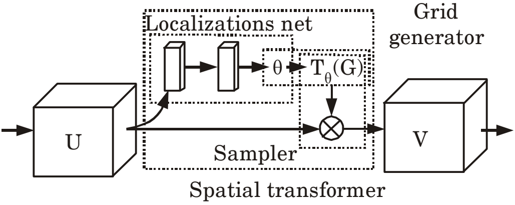


Fig. 4.7.1.

Que 4.8. **Define recurrent neural network.**

Answer

1. Recurrent Neural Network (RNN) are a type of neural network where the output from previous step are fed as input to the current step.
2. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words.
3. Thus RNN came into existence, which solved this issue with the help of a hidden layer.
4. The main and most important feature of RNN is hidden state, which remembers information about a sequence.
5. Recurrent neural networks are a family of neural networks for processing sequential data a recurrent neural network is a neural network that is specialized for processing a sequence of values $x^{(1)},\ldots, x^{(r)}$.
6. Most recurrent networks can also process sequences of variable length.
7. RNN has a memory which remembers all information about what has been calculated.
8. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output.
9. This reduces the complexity of parameters, unlike other neural networks.

Que 4.9. **What are the advantages and disadvantages of recurrent neural network ?**

Answer

Advantages of recurrent neural network :

1. An RNN remembers each and every information through time. It is useful in time series prediction only because of the feature to remember previous inputs as well. This is called Long Short Term Memory (LSTM).
2. Recurrent neural network are even used with convolutional layers to extend the effective pixel neighborhood.

Disadvantages of recurrent neural network :

1. Gradient vanishing and exploding problems.
2. Training an RNN is a very difficult task.
3. It cannot process very long sequences if using tanh or relu as an activation function.

Que 4.10. **Discuss the application of recurrent neural network.**

Answer

Applications of recurrent neural network are :

1. **Language modelling and prediction :**
   a. The probability of the output of a particular time-step is used to sample the words in the next iteration (memory).
   b. In language modelling, input is a sequence of words from the data and output will be a sequence of predicted word by the model.
   c. Output of the previous time step will be the input of the present time step.

2. **Speech recognition :**
   a. A set of inputs containing phonemes from an audio is used as an input.
   b. This network will compute the phonemes and produce a phonetic segment with the likelihood of output.

3. **Machine translation :**
   a. In machine translation, the input will be the source language (for example, Hindi) and the output will be in the target language (for example, English).
   b. The main difference between machine translation and language modelling is that the output starts only after the complete input has been fed into the network.

4. **Image recognition and characterization :**
   a. Recurrent neural network along with a convNet work together to recognize an image and give a description about it if it is unnamed.
   b. This combination of neural network works to produces fascinating results.

Que 4.11. **Explain working of RNN with the help of example.**

Answer

1. Suppose there is a deeper network with one input layer, three hidden layers and one output layer. Then like other neural networks, each hidden layer will have its own set of weights and biases, let's say, (w_1, b_1), (w_2, b_2) (w_3, b_3) are the weights and biases for hidden layer first, second and third respectively.
2. This means that each of these layers is independent of each other, *i.e.*, they do not memorize the previous outputs.

3. **Now RNN will perform the following task :**
   a. RNN converts the independent activations into dependent activations by providing the same weights and biases to all the layers, thus reducing the complexity of increasing parameters and memorizing each previous output by giving each output as input to the next hidden layer.
   b. Hence, these three layers can be joined together such that the weights and bias of all the hidden layers is the same, into a single recurrent layer.

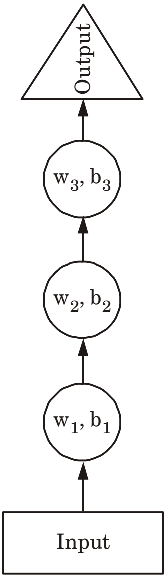


Fig. 4.11.1.

4. Formula for calculating current state :

$$h_t = f(h_{t-1}, x_t)$$

where,
h_t = Current state
h_{t-1} = Previous state
x_t = Input state

5. Formula for applying activation function (tanh) :

$$h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t)$$

where,
W_{hh} = Weight at recurrent neuron
W_{xh} = Weight at input neuron

6. Formula for calculating output :

$$y_t = W_{hy}h_t$$

$$y_t = \text{Output}$$
$$W_{hy} = \text{Weight at output layer}$$

Que 4.12. **Discuss LSTM.**

Answer

1. The idea of introducing self-loops to produce paths where the gradient can flow for long durations is a core contribution of the initial Long Short-Term Memory (LSTM) model.
2. A crucial addition has been to make the weight on this self-loop conditioned on the context, rather than fixed.
3. By making the weight of this self-loop gated (controlled by another hidden unit), the time scale of integration can be changed dynamically.
4. Integration can change based on the input sequence, because the time constants are output by the model itself.
5. The LSTM has been found extremely successful in many applications, such as unconstrained handwriting recognition, speech recognition, handwriting generation, machine translation, image captioning and parsing.
6. The forward propagation equations are given below, in the case of shallow recurrent network architecture.
7. Instead of a unit that simply applies an element wise nonlinearity to the affine transformation of input and recurrent units, LSTM recurrent networks have "LSTM cells" that have an internal recurrence (a self-loop), in addition to the outer recurrence of the RNN.
8. Each cell has the same inputs and outputs as an ordinary recurrent network, but has more parameters and a system of gating units that controls the flow of information.
9. The most important components is the state unit $s_i^{(t)}$ that has a linear self-loop similar to the leaky units. However, here, the self-loop weight is controlled by a forget gate unit $f_i^{(t)}$ (for time step t and cell i) that sets this weight to a value between 0 and 1 via a sigmoid unit :

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{ij}^f \, x_j^{(t)} + \sum_j W_{ij}^f \, h_j^{(t-1)}\right) \qquad \text{...(4.12.1)}$$

   where $x(t)$ is the current input vector and $h(t)$ is the current hidden layer vector, containing the outputs of all the LSTM cells, and b^f, U^f, W^f are respectively biases, input weights and recurrent weights for the forget gates.

10. The LSTM cell internal state is thus updated as follows, but with a conditional self-loop weight $f_i^{(t)}$:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{ij} x_j^{(t)} + \sum_j W_{ij} h_j^{(t-1)}\right)$$

...(4.12.2)

where b, U and W respectively denote the biases, input weights and recurrent weights into the LSTM cell.

11. The external input gate unit $g_i^{(t)}$ is computed similarly to the forget gate (with a sigmoid unit to obtain a gating value between 0 and 1), but with its own parameters :

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{ij}^g x_j^{(t)} + \sum_j W_{ij}^g h_j^{(t-1)}\right) \quad ...(4.12.3)$$

12. The output $h_i^{(t)}$ of the LSTM cell can also be shut off, via the output gate $q_i^{(t)}$, which also uses a sigmoid unit for gating :

$$h_i^{(t)} = \tan h\ (s_i^{(t)}) q_i^{(t)} \quad ...(4.12.4)$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{ij}^o x_j^{(t)} + \sum_j W_{ij}^o h_j^{(t-1)}\right) \quad ...(4.12.5)$$

which has parameters b^o, U^o, W^o for its biases, input weights and recurrent weights, respectively.

Que 4.13. **Discuss the training of network using RNN.**

Answer

Following are the steps involved in training of network using RNN :

1. A single time-step of the input is provided to the network.
2. Then calculate its current state using set of current input and the previous state.
3. The current h_t becomes h_{t-1} for the next time step.
4. One can go as many time steps according to the problem and join the information from all the previous states.
5. Once all the time steps are completed the final current state is used to calculate the output.
6. The output is then compared to the actual output *i.e.*, the target output and the error is generated.
7. The error is then back-propagated to the network to update the weights and hence the network (RNN) is trained.

Que 4.14. **Explain the Long Short-Term Memory Architecture.**

Answer

The chain-like architecture of LSTM allows to contain information for longer time periods, solving challenging tasks that traditional RNNs cannot solve.

The three major parts of the LSTM include :

1. **Forget gate :**

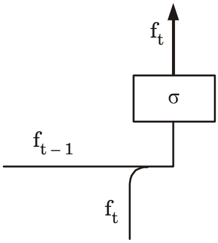


Fig. 4.14.1.

   a. A forget gate is responsible for removing information from the cell state.

   b. The information that is no longer required for the LSTM to understand things or the information that is of less importance is removed via multiplication of a filter.

   c. This is required for optimizing the performance of the LSTM network. This gate takes in two inputs, h_{t-1} and x_t.

   d. h_{t-1} is the hidden state from the previous cell or the output of the previous cell and x_t is the input at that particular time step.

   e. The given inputs are multiplied by the weight matrices and a bias is added.

2. **Input gate :**

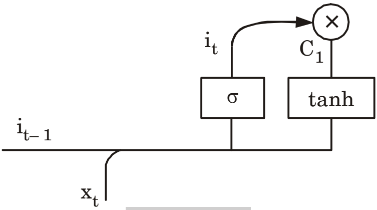


Fig. 4.14.2.

   a. The input gate is responsible for the addition of information to the cell state.

   b. This addition of information is three-step process as shown in Fig. 4.14.2.

1. Regulating what values need to be added to the cell state by involving a sigmoid function. This is similar to the forget gate and acts as a filter for all the information from h_{t-1} and x_t.

2. Creating a vector containing all possible values that can be added (as perceived from h_{t-1} and x_t) to the cell state. This is done using the tanh function, which outputs values from – 1 to + 1.

3. Multiplying the value of the regulatory filter (the sigmoid gate) to the created vector (the tanh function) and then adding this useful information to the cell state via addition operation.

3. Output gate :

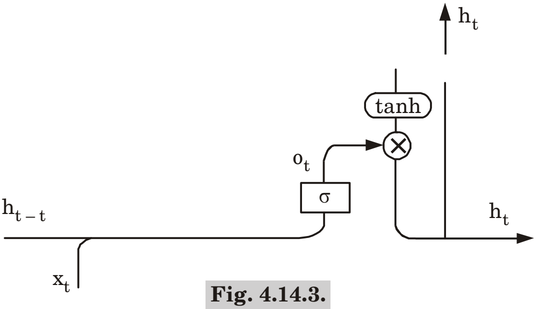


Fig. 4.14.3.

The functioning of an output gate can again be broken down to three steps :

1. Creating a vector after applying tanh function to the cell state, thereby scaling the values to the range – 1 to + 1.
2. Making a filter using the values of h_{t-1} and x_t, such that it can regulate the values that need to be output from the vector created above. This filter again employs a function.
3. Multiplying the value of this regulatory filter to the vector created in step 1, and sending it out as an output and also to the hidden state of the next cell.

Que 4.15. **What are the applications of LSTM ?**

Answer

Applications of LSTM include :

1. Language modelling :

a. A language model learns the probability of word occurrence based on examples of text.

b. Simpler models may look at a context of a short sequence of words, whereas larger models may work at the level of sentences or paragraphs.

c. Language models operate at the level of words.

2. Machine translation :

a. Machine translation is the task of automatically converting source text in one language to text in another language.

b. In a machine translation task, the input already consists of a sequence of symbols in some language, and the computer program must convert this into a sequence of symbols in another language.

3. **Image captioning :**
   a. Image captioning is the process of generating textual description of an image.
   b. It uses both natural language processing and computer vision to generate the captions.
4. **Question answer chatbots :**
   a. A Chatbot known as a conversational agent is a service either powered by rules or artificial intelligence that we interact via a chat interface.
   b. There are two main models for a chatbot :
      i. Retrieval-based model : This kind of chatbot uses a repository of predefined responses. The programmer chooses an appropriate response based on context following a given heuristic, which can be either something very simple or quite complex depending on the situation.
      ii. Generative model : A generative model chatbox does not use any predefined repository. This kind of chatbot is more advanced, because it learns from scratch using a process called deep learning.

Que 4.16. **What is recurrent neural networks language model ?**

Answer

1. Recurrent Neural Networks Language Model (RNNLM) is a type of neural networks language models which contains the RNNs in the network.
2. Since an RNN can deal with the variable length inputs, it is suitable for modelling the sequential data such as sentences in natural language.
3. We show one layer of an RNNLM with these parameters :

Symbol	**Definition**
x_t	The one-hot vector of t-th word
Y_t	The t-th output
$h_t^{(i)}$	The t-th hidden layer of i-th layer
p_t	The next word's probability of t-th word
E	Embedding matrix
W_h	Hidden layer matrix
W_o	Output layer matrix

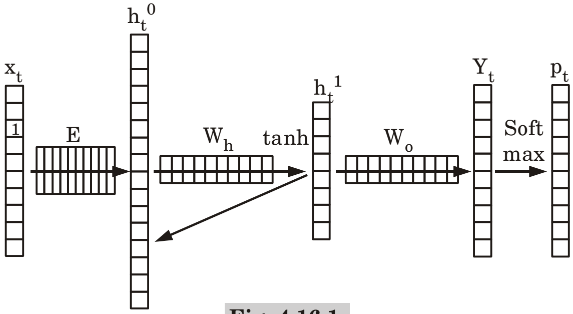


Fig. 4.16.1.

The process to get a next word prediction from i-th input word x_t

1. Get the embedding vector : $h_t^{(0)} = Ex_t$
2. Calculate the hidden layer : $h_t^{(1)} = \tanh\left(W_h\begin{bmatrix} h_t^{(0)} \\ h_{t-1}^{(1)} \end{bmatrix}\right)$
3. Calculate the output layer : $Y_t = W_o h_t^{(1)}$
4. Transform to probability : $p_t = \text{softmax}(Y_t)$

Que 4.17. **Discuss different types of recurrent neural network model.**

Answer

Different types of recurrent neural network languages models are :

1. ***n*-gram models :**
   a. In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech.
   b. The n-grams are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles.
   c. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of $(n - 1)$ order Markov model.
   d. n-gram models are used in probability, communication theory, computational linguistics (for instance, statistical natural language processing), computational biology (for instance, biological sequence analysis), and data compression.
   e. Two benefits of n-gram models are simplicity and scalability with larger n, a model can store more contexts with a well-understood space time trade off, enabling small experiments to scale up efficiently.

Que 4.18. **What are the applications of *n*-gram model ?**

Answer

***n*-gram model have been used to :**

1. Design kernels that allow machine learning algorithms such as support vector machines to learn from string data.
2. Find likely candidates for the correct spelling of a misspelled word.
3. Improve compression in compression algorithms where a small area of data requires *n*-grams of greater length.
4. Assess the probability of a given word sequence appearing in text of a language of interest in pattern recognition systems, speech recognition, OCR (Optical Character Recognition), Intelligent Character Recognition (ICR), machine translation.
5. Improve retrieval in information retrieval systems when it is hoped to find similar "documents" given a single query document and a database of reference documents.
6. Improve retrieval performance in genetic sequence analysis as in the BLAST family of programs.
7. Identify the language a text is in or the species a small sequence of DNA was taken from.
8. Predict letters or words at random in order to create text, as in the dissociated press algorithm.
9. It is used in cryptanalysis.

PART-3

Word-Level RNNs and Deep Reinforcement Learning, Computational and Artificial Neuroscience.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.19. **Define deep reinforcement learning.**

Answer

1. Deep reinforcement learning is a category of machine learning and artificial intelligence where intelligent machines can learn from their actions similar to the way humans learn from experience.

2. Inherent in this type of machine learning is that an agent is rewarded or penalised based on their actions. Actions that get them to the target outcome are rewarded (reinforced).
3. Through a series of trial and error, a machine keeps learning, making this technology ideal for dynamic environments that keep changing.
4. The deep portion of reinforcement learning refers to a multiple (deep) layers of artificial neural networks that replicate the structure of a human brain.
5. Deep learning requires large amounts of training data and significant computing power. Over the last few years, the volumes of data have exploded while the costs for computing power have reduced, which has enabled the explosion of deep learning applications.
6. The possibilities of deep reinforcement learning came to the attention of many during the well-publicised defeat of a Go grandmaster by DeepMind's AlphaGo.
7. In addition to playing Go, deep reinforcement learning has achieved human-level prowess in other games such as chess, poker, Atari games and several other competitive video games.

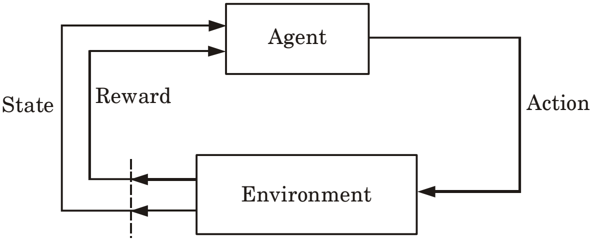


Fig. 4.19.1.

Que 4.20. **Write a short note on neuroscience.**

Answer

1. Neuroscience (or neurobiology) is the scientific study of the nervous system.
2. It is a multidisciplinary branch of biology that combines physiology, anatomy, molecular biology, developmental biology, cytology, mathematical modeling, and psychology to understand the fundamental and emergent properties of neurons and neural circuits.
3. The understanding of the biological basis of learning, memory, behavior, perception, and consciousness has been described as the "ultimate challenge" of the biological sciences.
4. Neurology works with diseases of the central and peripheral nervous systems, such as Amyotrophic Lateral Sclerosis (ALS) and stroke, and their medical treatment.

5. For example, brain imaging enables objective biological insight into mental illnesses, which can lead to faster diagnosis, more accurate prognosis, and improved monitoring of patient progress over time.
6. Integrative neuroscience describes the effort to combine models and information from multiple levels of research to develop a coherent model of the nervous system.
7. For example, brain imaging coupled with physiological numerical models and theories of fundamental mechanisms may shed light on psychiatric disorders.
8. The scope of neuroscience has broadened over time to include different approaches used to study the nervous system at different scales and the techniques used by neuroscientists have expanded enormously, from molecular and cellular studies of individual neurons to imaging of sensory, motor and cognitive tasks in the brain.

Que 4.21. **Explain computational neuroscience.**

Answer

1. Computational neuroscience is the field of study in which mathematical tools and theories are used to investigate brain function.
2. It can also incorporate diverse approaches from electrical engineering, computer science and physics in order to understand how the nervous system processes information.
3. Computational neuroscience is the only field that can help us to understand, how we are able to think and process information in our brain.
4. The ultimate goal of computational neuroscience is to explain how electrical and chemical signals are used in the brain to represent and process information.
5. It explains the biophysical mechanisms of computation in neurons, computer simulations of neural circuits, and models of learning.
6. Computational neuroscience is the theoretical study of the brain used to uncover the principles and mechanisms that guide the development, organization, information-processing and mental abilities of the nervous system.
7. Computational neuroscience is a specialization within neuroscience.

Que 4.22. **What do you mean by artificial neuroscience ?**

Answer

1. The Artificial Intelligence (AI) research field has presented a considerable growth in the last decades, helping researcher to explore new possibilities into their works.

2. Neuroscience's studies are characterized for recording high dimensional and complex brain data, making the data analysis computationally expensive and time consuming.
3. Neuroscience takes advantage of AI techniques and the increasing processing power in modern computers, which helped improving the understanding of brain behavior.
4. Artificial neuroscience has helped researchers to overcome the limitations raised when analyzing great amounts of data, making possible to explore new horizons in their areas.
5. Modern experimental methods in neuroscience areas such as brain imaging generate vast amount of high dimensional and complex data whose analysis represents a challenge Machine learning (ML) models without being explicitly programmed where to look are becoming ever more important for extracting reliable and meaningful relationships and for making accurate predictions.
6. ML models has been applied to analysis of neuropsychological data such as Magnetic Resonance Imaging (MRI), Near-Infrared Spectroscopy (NIRS), Electroencephalography (EEG), brain imaging, electromyography (EMG) as well as in a high-level.

Que 4.23. **What do you mean by natural language processing ? Why it is needed ?**

Answer

1. Natural language processing studies the problems inherent in the processing and manipulation of natural language and to make computer understand statements written in human language.
2. NLP can be defined as the automatic processing of human language.
3. Natural language processing is a subfield of AI which deals with the methods of communicating with a computer in one's own natural language.
4. It is used for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.
5. It is needed to bridge the gap between human and machine.
6. The goal of natural language is to enable people and computers to communicate in a natural language, such as English.
7. The field of NLP is divided into subfields :
   a. **NLU (Natural Language Understanding) :** This investigates methods of allowing the computer to comprehend instructions given in English.

b. **NLG (Natural Language Generation) :** This strive that computer produce ordinary English language so that people can understand computers more easily.

8. The study of language generation falls into following three areas :
   a. Determination of content.
   b. Formulating and developing a text utterance plan.
   c. Achieving a realization of the desired utterances.
9. A full NLU system would be able to :
   a. Paraphrase an input text.
   b. Translate the text into another language.
   c. Answer questions about the contents of the text.
   d. Draw inferences from the text.
11. Applications of NLP :
   a. Natural language interfaces to databases.
   b. Machine translation.
   c. Advanced word-processing tools.
   d. Explanation generation for expert systems.

Que 4.24. **What are the advantages and disadvantages of retrieval-based model and generative model ?**

Answer

Advantages of Retrieval-based model :

1. No grammatical or meaningless errors as we store the answers.
2. Works 100% well for the business problems and customer satisfaction and attention can be gained.
3. Easy to build these models as we do not require huge data.

Disadvantages of Retrieval-based model :

1. These systems do not generate any new text, they just pick a response from a fixed set.
2. A lot of hard coded rules have to be written.

Advantages of generative model :

1. No need to worry about the predefined responses and the rules.

Disadvantages of generative model :

1. Difficult to implement these and the output may not be accurate (grammatical / meaningless errors may occur).
2. Not applicable for the business problem (unless we are providing a service which may require text summarization techniques).
3. Huge data is required to train these models.

Que 4.25. **Compare recurrent neural network with multilayer neural network.**

Answer

S. No.	Recurrent neural network	Multilayer neural network
1.	Data and calculations flow in a backward direction, from the output data to the input.	Data and calculations flow in a single direction, from the input data to the outputs.
2.	It contains feedback links.	It does not contain feedback links.
3.	It is used for text data, speech data.	It is used for image data, time series of data.

Que 4.26. **Construct a recurrent network with four input nodes, three hidden nodes and four output nodes that has lateral inhibition structure in the output layer.**

Answer

A recurrent network with four input nodes, three hidden nodes and four output nodes are constructed as follows :

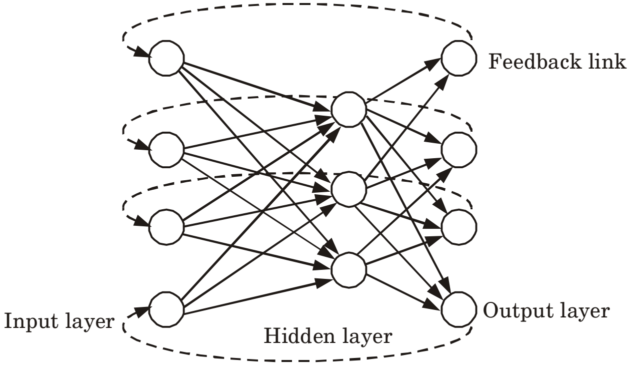


Fig. 4.26.1. A recurrent neural network with self loop.

☺☺☺

5 UNIT

Case Study and Applications

CONTENTS

PART-1

ImageNet, Detection, Audio WaveNet.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.1. **What is ImageNet ?**

Answer

1. ImageNet is a large database (dataset) of images which are originally labelled with Synsets of the WordNet lexicon tree. It was designed by academics intended for computer vision research.
2. It was the first of its kind in terms of scale. Images are organized and labelled in a hierarchy.
3. In Deep Neural Networks, machines are trained on a vast dataset of various images. Machines are required to learn useful features from these training images.
4. Once learned, they can use these features to classify images and perform many other tasks associated with computer vision.
5. ImageNet gives researchers a common set of images to benchmark their models and algorithms.
6. ImageNet is useful for computer vision applications such as object recognition, image classification and object localization.
7. ImageNet consists of 14,197,122 images organized into 21,841 sub-categories. These sub-categories can be considered as sub-trees of 27 high-level categories.
8. Thus, ImageNet is a well-organized hierarchy that makes it useful for supervised machine learning tasks.

Que 5.2. **Explain objectives of ImageNet with example.**

Answer

1. The ImageNet dataset contains images of fixed size of 224*224 and have RGB channels.
2. So, we have a tensor of (224, 224, 3) as input.
3. This model processes the input image and outputs a vector of 1000 values.

$$\hat{y} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_{999} \end{bmatrix}$$

4. This vector represents the classification probability for the corresponding class.
5. Suppose we have a model that predicts that image belongs to class 0 with probability 1, class 1 with probability 0.05, class 2 with probability 0.05, class 3 with probability 0.03, class 780 with probability 0.72, class 999 with probability 0.05 and all other class with 0, so, the classification vector for this will be :

$$\hat{y} = \begin{bmatrix} \hat{y}_0 = 0.1 \\ 0.05 \\ 0.05 \\ 0.03 \\ \vdots \\ \hat{y}_{780} = 0.72 \\ \vdots \\ \hat{y}_{999} = 0.05 \end{bmatrix}$$

6. To make sure these probabilities add to 1, we use softmax function. This is defined as :

$$P(y\,j \mid \Theta^{(i)}) = \frac{e^{\Theta^{(i)}}}{\sum_{j=0}^{K} e^{\Theta_K^{(i)}}}$$

7. After this we take the 5 most probable candidates into the vector.

$$C = \begin{bmatrix} 780 \\ 0 \\ 1 \\ 2 \\ 999 \end{bmatrix}$$

and our ground truth vector is defined as follows :

$$G = \begin{bmatrix} G_0 \\ G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} 780 \\ 2 \\ 999 \end{bmatrix}$$

8. Then we define our error function as follows :

$$E = \frac{1}{n} \sum_{K} \min_i d\,(C_i, G_K)$$

where $d = 0$ if $C_i = G_K$ else $d = 1$

9. So, the loss function for this example is :

$$E = \frac{1}{3}(\min_i d\,(C_i, G_1) + \min_i d\,(C_i, G_2) + \min_i d\,(C_i, G_3))$$

So, $E = \frac{1}{3}\ (0 + 0 + 0)$

$$E = 0$$

10. Since, all the categories in ground truth are in the predicted top 5 matrix, so the loss becomes 0.

Que 5.3. **Define the term object detection.**

Answer

1. Object detection is the act of finding the location of an object in an image.
2. Image classification labels the image as a whole. Finding the position of the object in addition to labeling the object is called object localization.
3. The position of the object is defined by rectangular coordinates.
4. For example, finding multiple objects in the image with rectangular coordinates is called detection.
5. The image shows four objects with bounding boxes. We will learn algorithms that can perform the task of finding the boxes.
6. The applications are enormous in robot vision, such as self-driving cars and industrial objects.
7. We can summarize localization and detection tasks to the following points :
   i. Localization detects one object in an image within a label.
   ii. Detection finds all the objects within the image along with the labels.
8. The difference is the number of objects. In detection, there are a variable number of objects.

9. This small difference makes a huge difference when designing the architectures for the deep learning model concerning localization or detection.

Que 5.4. **How to do object detection ? What is the problem with object detection ?**

Answer

Following are the steps taken to do object detection :

1. First, we take an image as input.
2. Then we divide the image into various regions.
3. We will then consider each region as a separate image.
4. Pass all these regions (images) to the CNN and classify them into various classes.
5. Once we have divided each region into its corresponding class, we can combine all these regions to get the original image with the detected object.

Problem with object detection :

1. The problem with using object detection approach is that the objects in the image can have different aspect ratios and spatial locations.
2. For instance, in some cases the object might be covering most of the image, while in others the object might only be covering a small percentage of the image.
3. The shapes of the objects might also be different (happens a lot in real-life use cases).
4. As a result of these factors, we would require a very large number of regions resulting in a huge amount of computational time.
5. So to solve this problem and reduce the number of regions, we can use region-based CNN, which selects the regions using a proposal method.

Que 5.5. **What is audio WaveNet in deep learning ? Explain.**

Answer

1. WaveNet is a deep neural network for generating raw audio.
2. The techniques able to generate realistic-sounding human-like voices by directly modelling waveforms using a neural network method trained with recordings of real speech.
3. WaveNet is an audio generative model based on the Pixel CNN. It is capable of producing audio that is very similar to a human voice.
4. There are experiments showing that WaveNet has improved current state-of-the-art Text-To-Speech (TTS) systems, reducing the difference with human voices by 50% for both US English and Mandarin Chinese.

5. In this generative model, each audio sample is conditioned on the previous audio sample. The conditional probability is modelled by a stack of convolutional layers.
6. This network does not have pooling layers, and the output of the model has the same time dimensionality as the input.

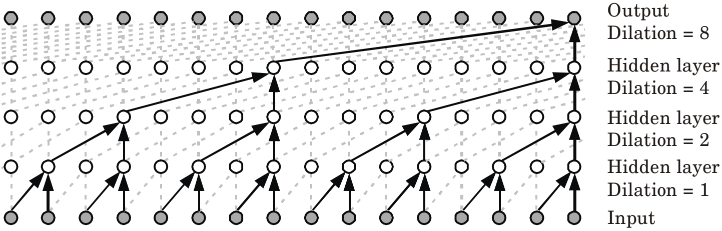


Fig. 5.5.1.

7. Fig. 5.5.1 shows how a WaveNet is structured. It is a fully convolutional neural network, where the convolutional layers have various dilation factors that allow its receptive field to grow exponentially with depth and cover thousands of time-steps.
8. At training time, the input sequences are real waveforms recorded from human speakers. After training, we can sample the network to generate synthetic utterances.
9. At each step during sampling a value is drawn from the probability distribution computed by the network. This value is then fed back into the input and a new prediction for the next step is made.

PART-2

Natural Language Processing Word2Vec, Joint Detection, Bioinformatics.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.6. **Discuss natural language processing Word2Vec in detail.**

Answer

1. Natural Language Processing (NLP) is an area of computer science and artificial intelligence that is known to be concerned with the interaction between computer and humans in natural language.

2. The goal is to enable the systems to fully understand various language as well as we do.
3. Word2Vec is one of the most widely used models to produce word embeddings.
4. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.
5. Word2Vec is a two-layer neural network that processes text. Its input is a text corpus and its output is a set of vectors (feature vectors for words in that corpus).
6. While Word2Vec is not a deep neural network, it turns text into a numerical form that deep networks can understand.
7. The purpose and usefulness of Word2Vec is to group the vectors of similar words together in vector space. That is, it detects similarities mathematically.
8. Word2Vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words. It does so without human intervention.
9. The output of the Word2Vec neural network is a vocabulary in which each item has a vector attached to it, which can be fed into a deep learning network or simply queried to detect relationships between words.
10. Word2Vec can be implemented in two ways, one is Skip Gram and other is Common Bag Of Words (CBOW).

Que 5.7. **Explain the steps involved in natural language processing.**

Answer

Steps involved in natural language processing :

Step 1 : Sentence segmentation :

a. Breaking the piece of text in various sentences.

Step 2 : Word tokenization :

a. Breaking the sentence into individual words called as tokens.
b. We can tokenize them whenever we encounter a space, we can train a model in that way.
c. Even punctuations are considered as individual tokens as they have some meaning.

Step 3 : Predicting parts of speech for each token :

a. Predicting whether the word is a noun, verb, adjective, adverb, pronoun, etc.
b. This will help to understand what the sentence is talking about.

c. This can be achieved by feeding the tokens (and the words around it) to a pre-trained part-of-speech classification model.

d. This model was fed a lot of English words with various parts of speech tagged to them so that it classifies the similar words it encounters in future in various parts of speech.

e. Again, the models do not really understand the 'sense' of the words, it just classifies them on the basis of its previous experience. It is pure statistics.

Step 4 : Lemmatization :

a. Feeding the model with the root word.

Step 5 : Identifying stop words :

a. There are various words in the English language that are used very frequently like '*a*', and, 'the' etc.

b. These words make a lot of noise while doing statistical analysis. We can take these words out. Some NLP pipelines will categorize these words as stop words, they will be filtered out while doing some statistical analysis.

c. Definitely, they are needed to understand the dependency between various tokens to get the exact sense of the sentence.

d. The list of stop words vary and depends on what kind of output are we expecting.

Step 6 : Dependency parsing :

a. This means finding out the relationship between the words in the sentence and how they are related to each other.

b. We create a parse tree in dependency parsing, with root as the main verb in the sentence.

Step 7 : Finding noun phrases :

a. We can group the words that represent the same idea.

Step 8 : Named Entity Recognition (NER) :

a. NER maps the words with the real world places.

b. The places that actually exist in the physical world.

c. We can automatically extract the real world places present in the document using NLP.

Que 5.8. **Explain Langmod_nn model with its layer.**

Answer

Langmod_nn model :

1. The Langmod_nn model builds a three-layer Forward Bigram Model neural network where the goal is to use a given word in a corpus to attempt to predict the next word.

2. The model consists of the following three layers :

a. **Embedding layer :**

   i. Each word corresponds to a unique embedding vector, a representation of the word in some embedding space.

   ii. We find the embedding for a given word by doing a matrix multiply (essentially a table lookup) with an embedding matrix that is trained during regular backpropagation.

b. **Hidden layer :** A fully-connected feed-forward layer with hidden layer size 100, and Rectified Linear Unit (ReLU) activation.

c. **Softmax layer :**

   i. A fully-connected feed-forward layer with layer size equal to the vocabulary size, where each element of the output vector corresponds to the probability of that word in the vocabulary being the next word.

Que 5.9. **What are the applications of natural language processing ?**

Answer

Following are the applications of natural language processing :

1. **Healthcare :** A healthcare solution by Nuance, Dragon Medical One is capable of allowing doctors to dictate basic medical history, progress notes and even future plans of action directly into their EHR.

2. **Computerized personal assistants and personal virtual assistance :**

   a. It is a known fact that one of NLP's largest application in the modern era has been in the design of personal voice assistants like Siri, Cortana and Alexa.

   b. But imagine being able to tell Siri to set up a meeting with your boss. Imagine if then, Siri was capable of somehow comparing your schedule to that of your boss, being able to find a convenient time for your meeting and then revert back to you and your boss with a meeting all fixed. This is what is called a Personal Virtual Assistant (PVA).

3. **Customer service :**

   a. Using advanced concepts of natural language processing, it might be possible to completely automate the process of handling customers that call into call centers.

   b. Not only this, it might become easier to retrieve data from an unorganized structure for said customers using such a solution.

4. **Sentiment analysis :**

   a. NLP has been used extensively to determine the sentiment behind the tweets/posts of users that take to the internet to share their emotions.

b. Not only that, it may be possible to use sentiment analysis to detect depression and suicidal tendencies.

Que 5.10. **Explain different types of algorithm used in NLP.**

Answer

Different types of algorithm used in NLP are :

1. **Naive Bayes algorithm :** The Naive Bayesian Analysis (NBA) is a classification algorithm that is based on the Bayesian Theorem, with the hypothesis on the feature's independence.

$$P(C \mid x) = \frac{P(C \mid x) \times P(C)}{P(x)}$$

2. **Perceptron :** Refer Q. 1.8, Page 1–7M, Unit-1.
3. **Support vector machine :** Refer Q. 1.7, Page 1–6M, Unit-1.
4. **Logistics regression :** Refer Q. 1.9, Page 1–9M, Unit-1.

Que 5.11. **Explain advantages and disadvantages of different types of NLP.**

Answer

1. **Naive Bayes :**

Advantages :

1. Easy to implement.
2. Estimation is fast, requiring only a single pass over the data.
3. Assigns probabilities to predicted label.
4. Controls overfitting with smoothing parameter.

Disadvantage :

i. Often has poor accuracy, especially with correlated features.

2. **Perceptron :**

Advantages :

1. Easy to implement.
2. Error-driven learning means that accuracy is typically high, especially after averaging.

Disadvantage :

1. Not probabilistic.
2. Hard to know when to stop learning.
3. Lack of margin can lead to overfitting.

3. Support vector machine :

Advantages :

1. Optimizes an error-based metric.
2. Usually resulting in high accuracy.
3. Overfitting is controlled by a regularization parameter.

Disadvantage : Not probabilistic.

4. Logistic regression :

Advantages :

1. Error driven and probabilistic.
2. Overfitting is controlled by a regularization parameter.

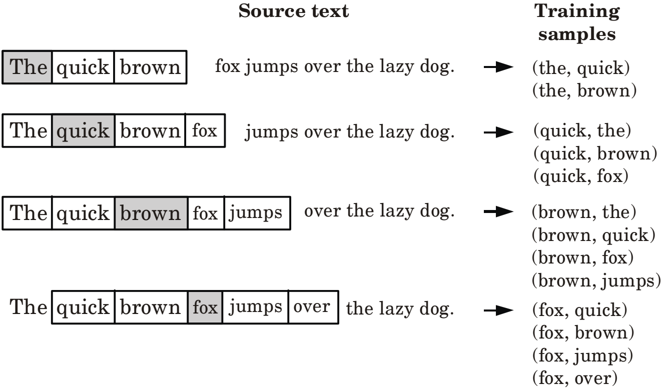
Disadvantages :

1. Batch learning requires black-box optimization.
2. Logistic loss can "overtrain" on correctly labeled examples.

Que 5.12. Explain Skip Gram and CBOW versions of Word2Vec.

Answer

1. CBOW is learning to predict the word by the context.
2. Here the input will be the context neighboring words and output will be the target word.
3. The limit on the number of words in each context is determined by a parameter called 'window size'.



4. The quick brown fox jumps over the lazy dog :

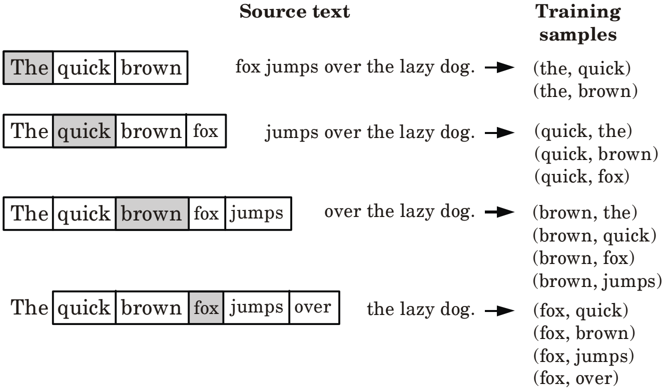
   Model : CBOW

   Input layer : White box content

   Target layer : blue box word

   Window size : 5

5. Skip gram is learning to predict the context by the word. Here the input will be the word and output will be the target context neighboring words. The limit on the number of words in each context is determined by a parameter called "window size".



6. The quick brown fox jumps over the lazy dog :

Model : Skip Gram

Input layer : blue box word

Target layer : White box content

Window size : 5

Que 5.13. **Define Bioinformatics.**

Answer

1. Bioinformatics is a field of study that uses computation to extract knowledge from biological data.
2. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software.
3. Bioinformatics deals with computational and mathematical approaches for understanding and processing biological data.
4. It is an interdisciplinary field in which new computational methods are developed to analyze biological data and to make biological discoveries.
5. Bioinformatics is considered to be a much broader discipline, encompassing modelling and image analysis in addition to the classical methods used for comparison of linear sequences or three-dimensional structures.
6. Application of machine learning in bioinformatics has given rise to a lot of application from diseases prediction, diagnosis and survival analysis.

7. Data science has changed a lot in bioinformatics from dimensionality reduction of large datasets to data visualisation.

Machine Learning in Biology :

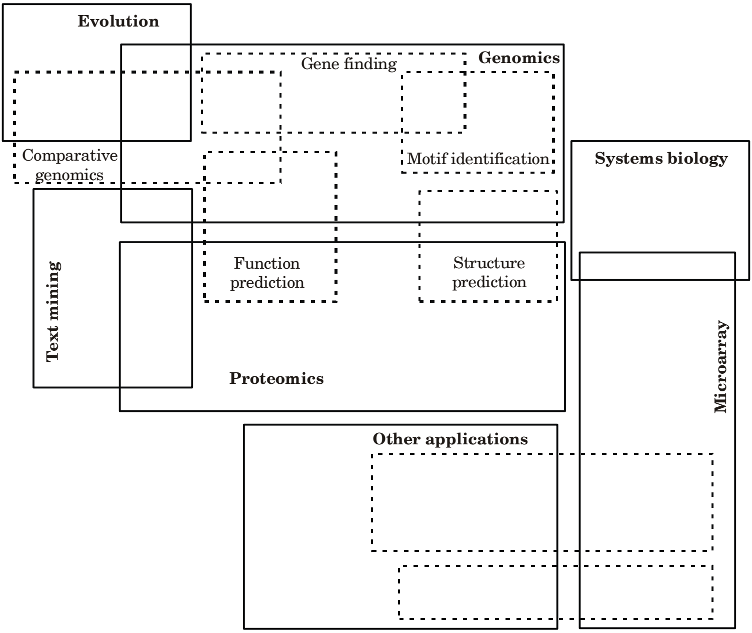


Fig. 5.13.1.

1. Presently a large list of bioinformatics tools and software are available which are based on machine learning.
2. The twin of bioinformatics, called computational biology have emerged into development of software and application using machine learning and deep learning techniques for biological image data analysis.
3. Google's Deep Learning library called TensorFlow was shown how it can be used in computational biology.
4. Application of machine learning and deep learning in biology need to be explored further for building AI's which can be used for disease diagnosis and prediction.

Que 5.14. **Explain Bioinformatics tools.**

Answer

Major category of Bioinformatics tools are :

1. **Homology and similarity tools :**

a. The concept of homology common evolutionary ancestry is central to computational analyses of protein and DNA sequences, but the link between similarity and homology is often misunderstood.

b. We infer homology when two sequences or structures share more similarity than would be expected by chance when excess similarity is observed, the simplest explanation for that excess is that the two sequences did not arise independently, they arose from a common ancestor.

c. Common ancestry explains excess similarity (other explanations require similar structures to arise independently) thus excess similarity implies common ancestry.

2. **Protein function analysis :**

a. Proteins differ from each other in their size, molecular structure and physiochemical properties.

b. These differences allow for protein analysis and characterization by separation and identification. Separation is done via electrophoresis where proteins are differentiated by size or mass, and isoelectric focusing, where protein are separated by charge.

c. These techniques can be done independently, or in combination referred to as 2D electrophoresis. Equipment for separation includes polyacrylamide gels, organic stains, electrophoresis boxes, isoelectric focusing immobilized strips, 2D electrophoresis equipment, and protein standards.

d. Identification is done via mass spectrometry where molecules are ionized to determine their mass to charge ratios.

3. **Structural analysis :**

a. Structural analysis is a fundamental tool to better evaluate the seismic response and vulnerability of historical buildings and define effective strengthening interventions.

b. In particular, the use of advanced numerical tools to perform three-dimensional (3D) nonlinear dynamic analyses allows obtaining a thorough detailed knowledge of the seismic behavior of such a typology of structures.

4. **Sequence analysis :**

a. In bioinformatics, sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution.

b. Methodologies used include sequence alignment, searches against biological databases, and others.

Que 5.15. **Discuss the use of Bioinformatics.**

Answer

Use of Bioinformatics :

1. **Molecular modelling :**
   - i. In 3D structure prediction
   - ii. In Protein function prediction
2. **Molecular interactions :**
   - i. In Protein-protein docking
   - ii. Finding inhibitors, activators of proteins
   - iii. In Protein-DNA interactions
3. **Phylogenetic analyses :**
   - i. In Re-construction of evolution history
   - ii. In Tracking gene flow
   - iii. In Identification of conserved regions
4. **Protein sequence analyses :**
   - i. In Molecular mass, instability index, amino-acid composition
   - ii. In Signals peptide identification
   - iii. In Secondary structure analyses
5. **Drug designing :**
   - i. In Target identification
   - ii. In Target validation
   - iii. In Lead identification

Que 5.16. **Discuss various branches of bioinformatics.**

Answer

Various branches of bioinformatics are :

1. **Animal bioinformatics :**
   - a. It deals with computer added study of genomics, proteomics and metabolomics in various animal species.
   - b. It includes study of gene mapping, gene sequencing, animal breeds, animal genetic resources etc.
   - c. It can be further divided as bioinformatics of mammal's reptiles, insects, birds, fishes etc.
2. **Plant bioinformatics :**
   - a. It deals with computer aided study of plant species.
   - b. It includes gene mapping, gene sequencing, plant genetic resources, database etc.

c. It can be further divided into following branches :

   i. **Agricultural bioinformatics :** It deals with computer based study of various agricultural crop species. It is also referred to as crop bioinformatics.

   ii. **Horticultural bioinformatics :** It refers to computer aided study of horticultural crops, viz., fruit crops, vegetable crops and flower crops.

   iii. **Medicinal plants bioinformatics :** It deals with computer based study of various medicinal plant species.

   iv. **Forest plant bioinformatics :** It deals with computer based study of forest plant species.

Que 5.17. **What are the advantages and disadvantages of bioinformatics ?**

Answer

Advantages of bioinformatics :

1. It provides systematic information about genomics, proteomics and metabolomics of living organisms. This information is useful in planning various breeding and genetical programmes.
2. It helps in finding evolutionary relationship between two species. Studies of nucleotide and protein sequences help in such matter. The closely related organisms have similar sequences and distantly related organisms have dissimilar sequence.
3. It is a rapid method of gene mapping and sequencing.
4. Computer aided studies help in identification of similar genes in two species.
5. The computer based information has very high level of accuracy and is highly reliable.
6. Bioinformatics has led to advances in understanding basic biological processes which in turn have helped in diagnosis, treatment and prevention of many genetic diseases.
7. It has become possible to reconstruct genes from Expressed Sequence Tags (EST). The EST is a short piece of genes which can express.
8. Computer aided programmes have made it possible to group proteins into families based on their relatedness.
9. Computer aided programmes are useful in designing primers for PCR.
10. In life science, computer aided programmes are useful in storing, organizing and indexing huge databases.

Disadvantages of bioinformatics :

1. Bioinformatics requires sophisticated laboratory of molecular biology for in-depth study of biomolecules. Establishment of such laboratories requires lot of funds.

2. Computer based study of life science requires training about various computer programmes applicable for the study of different processes of life science. Thus special training is required for handling of computer based biological data.
3. There should be uninterrupted electricity (power) supply for computer aided biological investigations. Interruption of power may lead to loss of huge data from the computer memory.
4. There should be regular checking of computer viruses because viruses may pose several problems such as deletion of data and corruption of the programmes.
5. The maintenance and up keeping of molecular laboratories involves lot of expenditure which becomes a limiting factor for computer based molecular studies.

PART-3

Face Recognition, Scene Understanding, Gathering Image Captions.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.18. **Discuss face recognition in deep learning.**

Answer

1. Face recognition is a method of identifying or verifying the identity of an individual using their face. Face recognition systems can be used to identify people in photos, video, or in real-time.
2. Face recognition systems use computer algorithms to pick out specific, distinctive details about a person's face.
3. These details, such as distance between the eyes or shape of the chin, are then converted into a mathematical representation and compared to data on other faces collected in a face recognition database.
4. The data about a particular face is often called a face template and is distinct from a photograph because it is designed to only include certain details that can be used to distinguish one face from another.
5. Face recognition is often described as a process that first involves four steps :
   a. **Face detection :** Locate one or more faces in the image and mark with a bounding box.

b. **Face alignment :** Normalize the face to be consistent with the database, such as geometry and photometrics.

3. **Feature extraction :** Extract features from the face that can be used for the recognition task.

4. **Face recognition :** Perform matching of the face against one or more known faces in a prepared database.

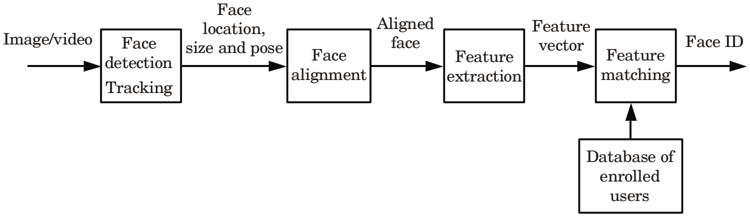


Fig. 5.18.1. Face recognition processing flow.

Que 5.19. **Explain the steps used in face recognition.**

Answer

Following are the steps used in face recognition :

Step 1 : A picture of our face is captured from a photo or video. Our face might appear alone or in a crowd. Our image may show us looking straight ahead or nearly in profile.

Step 2 : Facial recognition software reads the geometry of our face. Key factors include the distance between our eyes and the distance from forehead to chin. The software identifies facial landmarks - one system identifies 68 of them - that are keys to distinguishing our face. The result: our facial signature.

Step 3 : Our facial signature a mathematical formula is compared to a database of known faces.

Step 4 : A determination is made. Our faceprint may match that of an image in a facial recognition system database.

Que 5.20. **Discuss the application of face (facial) recognition.**

Answer

Application of facial recognition :

1. **US government at airports :**

   a. Facial recognition systems can monitor people coming and going in airports.

   b. The department of homeland security has used the technology to identify people who have overstayed their visas or may be under criminal investigation.

2. **Mobile phone makers in products :**
    a. Apple first used facial recognition to unlock its iPhone X, and continues with the iPhone XS.
    b. Face ID authenticates and makes sure we are accessing phone.
3. **Colleges in the classroom :** Facial recognition software can, in essence, take roll. If we decide to miss class, our professor could know.
4. **Social media companies on websites :**
    a. Facebook uses an algorithm to spot faces when we upload a photo to its platform.
    b. The social media company asks if want to tag people in our photos. If we say yes, it creates a link to their profiles.
    c. Facebook can recognize faces with 98 percent accuracy.
5. **Businesses at entrances and restricted areas :** Some companies have traded in security badges for facial recognition systems.
6. **Religious groups at places of worship :** Churches have used facial recognition to scan their congregations to see who's present.
7. **Retailers in stores :** Retailers can combine surveillance cameras and facial recognition to scan the faces of shoppers. One goal is to identifying suspicious characters and potential shoplifters.
8. **Airlines at departure gates :** When the fliers board the flight the airline scans the flier face.

Que 5.21. **What are the issues related with face recognition ?**

Answer

Issues related with face recognition :

1. **Security :** Our facial data can be collected and stored, often without our permission. It's possible hackers could access and steal that data.
2. **Prevalence :** Facial recognition technology is becoming more widespread. That means our facial signature could end up in a lot of places. We probably would not know who has access to it.
3. **Ownership :** We own our face the one atop our neck but our digital images are different. We may have given up our right to ownership when we signed up on a social media network.
4. **Safety :** Facial recognition could lead to online harassment and stalking. For example, someone takes our picture on a subway or some other public place and uses facial recognition software to find out exactly who we are.
5. **Mistaken identity :** Facial recognition systems may not be 100 percent accurate.
6. **Basic freedoms :** Government agencies and others could have the ability to track us. What we do and where we go might no longer be private. It could become impossible to remain anonymous.

Que 5.22. **How is deep learning used in facial recognition ?**

Answer

1. Deep learning networks are loosely based on the structure of the human brain, and enable us to train machines to learn by example.
2. This means that once the deep learning algorithms have been trained for long enough using datasets that are both sufficiently large and diverse, they can apply what they have learned to make predictions or produce results in response to new data.
3. Deep learning in the form of Convolutional Neural Networks (CNNs) to perform facial recognition.
4. A CNN is a type of Deep Neural Network (DNN) that is optimized for complex tasks such as image processing, which is required for facial recognition.
5. CNNs consist of multiple layers of connected neurons. There is an input layer, an output layer, and multiple layers between these two.
6. With facial recognition, the input is an image, which the CNN processes as groups of pixels. These groups are scanned as matrices, and the values within the matrices are multiplied, with the results of this multiplication being fed into the next layer.
7. This process continues through all the layers, until it reaches the output layer, where the network produces an output in the form of an array of 2048 numbers. This array is referred to as a faceprint.
8. The computed faceprint can then be compared to another faceprint (1:1 matching), or to a database of faceprints (1:N matching), to determine whether or not there is a match.
9. If two or more faceprints are similar enough, based on the chosen confidence thresholds, they will be recorded as a match.

Que 5.23. **Write short note on face detection.**

Answer

1. Face detection is a fundamental step in facial recognition and verification.
2. It also extends to a broad range of other applications including facial expression recognition, face tracking for surveillance purposes, digital tagging on social media platforms and consumer applications in digital technologies, such as auto-focusing ability in phone cameras.
3. This survey will examine facial detection methods as applied to facial recognition and verification, the greatest obstacle faced by face detection algorithms was the ability to achieve high accuracy in uncontrolled conditions.

4. Their usability in real life applications was limited.
5. However, since the development of the viola jones boosting based face detection method face detection in real life has become commonplace.
6. Significant progress has since been made by researchers in this area due to the development of powerful feature extraction techniques including Scale Invariant Feature Transform (SIFT).
7. Histograms of oriented Gradients (HoGs).
8. Local Binary Patterns (LBPs) and methods such as Integral Channel Features (IVF).
9. For a recent and comprehensive review of these traditional face detection methodologies, readers are referred.
10. This review will focus on more recently proposed deep learning methods, which were developed in response to the limitations of HoG and Haar wavelet features in capturing salient facial information under unconstrained conditions which include large variations in resolutions, illumination, pose, expression, and color.

Que 5.24. **What is scene understanding ?**

Answer

1. Scene understanding is the process of perceiving, analysing and elaborating an interpretation of a 3D dynamic scene observed through a network of sensors.
2. This process consists in matching signal information coming from sensors observing the scene with models which humans are using to understand the scene.
3. Based on that, scene understanding is both adding and extracting semantic from the sensor data characterizing a scene.
4. This scene can contain a number of physical objects of various types (for example, people, vehicle) interacting with each other or with their environment (for example, equipment) more or less structured.
5. The goal of scene understanding is to obtain as much semantic knowledge of a given scene image as possible. This include categorization (labelling the whole scene), object detection (predicting object locations by bounding boxes), and semantic segmentation (labelling each pixel).
6. Due to this very general formulation, there is a wide range of applications, such as urban scene understanding for automotive applications, generic object detection, or inferring semantics of remote sensing data.
7. Scene understanding can achieve four levels of generic computer vision functionality of detection, localisation, recognition and understanding.
8. The key characteristic of a scene understanding system is its capacity to exhibit robust performance even in circumstances that were not foreseen when it was designed.

9. Most of the works, which use deep learning for integration, are based on the combination of depth information and semantic segmentation.
10. The use of deep learning to combine the tasks of text detection, object recognition, scene classification and caption generation remains an open research field.
11. The need is to integrate these components into a combined framework to aid in the development of a low cost and robust scene understanding system.

Que 5.25. **Explain scene classification.**

Answer

1. **Scene classification :**
   a. The goal in this experiment is to classify an unknown image as one of the eight learned scene classes.
   b. We perform three experiments to analyze the different aspects of our model and learning approach.
   c. All evaluations are done based on the 8-way classification results.
2. **Comparison with different models :**
   a. We compare the results of our model with three other approaches :
      i. A baseline bag of words image classification model.
      ii. The region-based model used to initialize our initial object class models.
      iii. A modified Corr-LDA model based on by adding a class variable on top of the mixing proportation parameter θ in the original model.
3. **Influence of unannotated data :**
   a. To provide some in-sight into the learning process.
   b. Left the classification performance curve as a function of the number of unlabeled images given to the model.
   c. In this experiment, the number of initialized images is fixed to 30.
   d. Performance gradually increases when more unlabeled images are included.
   e. This proves the effectiveness of unlabeled data in our learning framework.
4. **Effect of noise in tags :**
   a. In order to underline the robustness of our model to noisy training data, we present a set of experiments in which we dilute the original flickr tags with different percentages of noise by adding arbitrary words from the list of 1256 words during the training process.

b. Right shows that while the algorithm of decreases in accuracy when noise increases, our model is oblivious to even large percentages of noise.

c. The robustness to noise is mostly attributed to the switch variable the image without any such information given during training.

d. We first compare quantitative results with another approach and then show a qualitative difference in example segmentations with and without the top down contextual influence provided by the scene class C.

5. **Comparison to another segmentation method :**

a. In the image segmentation and annotation experiments, we train our model on 30 initialized images plus 170 unlabeled images.

b. We test on 240 images where ground truth is provided by human segmentation. Precision is computed by dividing the total area of correctly segmented pixels by the total area of detected pixels for each object.

c. Recall is calculated by dividing the total area of correctly segmented pixels by the total area of true pixels of each object.

d. We compare our segmentation results with the region-based model used in the training of our initial object models.

e. It is also one of the state-of-the-art concurrent object segmentation and recognition methods. Table shows that our model significantly outperforms in every object classes.

6. **Influence of the scene class on annotation and segmentation :**

a. In this experiment, we examine the top-down, contextual influence of a scene in our model.

b. We compare our full model to a damaged model in which the top down influence of the scene class is ignored.

c. Our results underscore the effectiveness of the contextual facilitation by the top-down classification on the annotation and segmentation tasks.

Que 5.26. **Explain gathering image captions.**

Answer

1. Image captioning is the process of generating textual description of an image. It uses both natural language processing and computer vision to generate the captions.

2. The dataset will be in the form [image → captions]. The dataset consists of input images and their corresponding output captions.

3. Image captioning refers to the process of generating textual description from an image based on the objects and actions in the image.

4. The task of image captioning can be divided into two modules logically one is an image-based model which extracts the features and nuances out of our image, and the other is a language based model which translates the features and objects given by our image based model to a natural sentence.

5. For our image-based model (viz encoder), we rely on a Convolutional Neural Network model. For our language-based model (viz decoder), we rely on a Recurrent Neural Network.

Que 5.27. **What is computer vision ? What are the types of computer vision ?**

Answer

Computer vision is a field of artificial intelligence that trains computers to interpret and understand the visual world. Using digital images from cameras and videos and deep learning models, machines can accurately identify and classify objects and then react to what they see.

Different types of computer vision :

1. **Image segmentation :** It partitions an image into multiple regions or pieces to be examined separately.
2. **Object detection :** It identifies a specific object in an image. Advanced object detection recognizes many objects in a single image; a football field, an offensive player, a defensive player, a ball and so on. These models use an *X*, *Y* coordinate to create a bounding box and identify everything inside the box.
3. **Facial recognition :** It is an advanced type of object detection that not only recognizes a human face in an image, but identifies a specific individual.
4. **Edge detection :** It is a technique used to identify the outside edge of an object or landscape to better identify what is in the image.
5. **Pattern detection :** It is a process of recognizing repeated shapes, colors and other visual indicators in images.
6. **Image classification :** It groups images into different categories.
7. **Feature matching :** It is a type of pattern detection that matches similarities in images to help classify them.

☺☺☺

1 UNIT

Introduction to Machine Learning (2 Marks Questions)

1.1. Define activation function.

Ans. An activation function is the basic element in neural model. It is used for limiting the amplitude of the output of a neuron. It is also called squashing function.

1.2. What are different types of activation functions ?

Ans. **Types of activation function :**

1. Signum function
2. Sigmoidal function
3. Identity function
4. Binary step function
5. Bipolar step function

1.3. Give the difference between supervised and unsupervised learning in artificial neural network.

Ans.

S. No.	Supervised learning	Unsupervised learning
1.	It uses known and labeled data as input.	It uses unknown data as input.
2.	It uses offline analysis.	It uses real time analysis of data.
3.	Number of classes are known.	Number of classes are not known.
4.	Accurate and reliable results.	Moderate accurate and reliable results.

1.4. What are neurons ?

Ans. A neuron is a small cell that receives electro chemical signals from its various sources and in return responds by transmitting electrical impulses to other neurons.

1.5. Give advantages of neural network.

Ans. **Advantages of neural network :**

1. A neural network can perform tasks that a linear program cannot perform.
2. It can be implemented in any application.
3. A neural network does not need to be reprogrammed.

1.6. What are the disadvantages of neural network ?

Ans. **Disadvantages of neural network :**

1. The neural network needs training to operate.
2. It requires high processing time for large neural network.

1.7. What is single layer feedforward network ?

Ans. Single layer feedforward network is the simplest form of a layered network where an input layer of source nodes that projects onto an output layer of neurons, but not vice versa.

1.8. Write different applications of Neural Networks (NN).

Ans. **Applications of NN are :**

1. Image recognition
2. Data mining
3. Machine translation
4. Spell checking
5. Stock prediction
6. Statistical modeling

1.9. Why sigmoid function is so important activation function in neural network ?

Ans. A sigmoid function takes any real number as input and returns a value between 0 and 1. Therefore, it is especially used for models where we have to predict the probability as an output. Since probability of anything exists only between the range of 0 and 1, so sigmoid function is important in neural network.

1.10. What is gradient descent ?

Ans. Gradient descent is an optimization technique in machine learning and deep learning and it can be used with all the learning algorithms. A gradient is the slope of a function, the degree of change of a parameter with the amount of change in another parameter.

1.11. What are the types of gradient descent ?

Ans. **Types of gradient descent :**

1. Batch gradient descent
2. Stochastic gradient descent
3. Mini-batch gradient descent

1.12. What is backpropagation algorithm ?

Ans. Backpropagation is an algorithm used in the training of feedforward neural networks for supervised learning. Backpropagation efficiently computes the gradient of the loss function with respect to the weights of the network for a single input-output example.

1.13. What is logistic regression ?

Ans. Logistic regression is a supervised classification algorithm. It is based on maximum likelihood estimation. In a classification problem, the target variable (output) y, can take only discrete values for given set of features (inputs) x.

1.14. What are the issues related with machine learning ?

Ans. **Issues related with machine learning :**

1. Data quality
2. Transparency
3. Manpower

1.15. What is loss function ?

Ans. Loss function estimates how well particular algorithm models the provided data.

1.16. What are the classes of problems in machine learning ?

Ans. **Common classes of problem :**

1. Classification
2. Regression
3. Clustering
4. Rule extraction

1.17. What are the two kinds of signals that are identified in multilayer perceptron ?

Ans. **Two kinds of signals that are identified in multilayer perceptron :**

1. Functional signal
2. Error signal.

1.18. List the two classes of loss function.

Ans. **Two classes of loss function :**

1. Regression losses
2. Classification losses.

1.19. List the tunning parameter that effects the backpropogation neural network.

Ans. **Tunning parameter that effects the backpropogation neural networks :**

1. Momentum factor

2. Learning coefficient
3. Sigmoidal gain
4. Threshold value

1.20. What are the various parameters in backpropogation neural network ?

Ans. Various parameters in backpropogation neural networks are :

1. Number of hidden nodes
2. Momentum coefficient α
3. Sigmoidal gain λ
4. Local minima

☺☺☺

2 UNIT

Deep Networks (2 Marks Questions)

2.1. What is deep learning ?

Ans. Deep learning is the subfield of artificial intelligence that focuses on creating large neural network models that are capable of making accurate data-driven decisions. Deep learning is used where the data is complex and large datasets available.

2.2. What are different types of deep learning algorithm ?

Ans. Different types of deep learning algorithm :

1. Feedforward neural networks
2. Radial basis function neural network
3. Multilayer perceptron
4. Unsupervised pre-trained network
5. Convolutional neural networks
6. Recurrent neural network
7. Recursive neural networks

2.3. Define the term deep neural networks.

Ans. Deep Neural Networks (DNNs) are composed of multiple levels of non-linear operations, such as neural networks with many hidden layers. Deep learning methods aim at learning feature hierarchies, where features at higher levels of the hierarchy are formed using the features at lower levels.

2.4. What are the applications of semi-supervised learning ?

Ans. Applications of semi-supervised learning :

1. Speech analysis
2. Internet content classification
3. Protein sequence classification

2.5. Why is semi-supervised learning important ?

Ans. When we do not have enough labeled data to produce an accurate model and we do not have the ability or resources to get more data, we can use semi-supervised techniques to increase the size of our training data.

2.6. What do you mean by generalization ?

Ans. The ability to perform well on previously unobserved inputs is called generalization.

2.7. What are the factors that determine how well a machine learning algorithm will perform?

Ans. The factors determining how well a machine learning algorithm will perform is its ability to :

1. Make the training error small.
2. Make the gap between training and test error small.

2.8. How does underfitting and overfitting occurs ?

Ans. Underfitting occurs when the model is not able to obtain a sufficiently low error value on the training set. Overfitting occurs when the gap between the training error and test error is too large.

2.9. What are the primary tasks of convolutional neural networks ?

Ans. The primary tasks of convolutional neural networks are :

1. Classify visual content (describe what they "see").
2. Recognize objects within is scenery (for example, eyes, nose, lips, ears on the face).
3. Gather recognized objects into clusters (for example, eyes with eyes, noses with noses).

2.10. Name the layers used in convolutional neural network.

Ans. Layers used in convolutional neural networks are :

1. Convolutional layer
2. Rectified Linear Unit layer
3. Pooling layer
4. Connected layer

2.11. Give some applications of CNN.

Ans. Application of CNN are :

1. Decoding facial recognition
2. Analyzing documents
3. Understanding climate
4. Historic and environmental collections
5. Grey areas
6. Advertising

2.12. What are the applications of deep learning ?

Ans. Applications of deep learning :

1. Automatic text generation
2. Healthcare
3. Automatic machine translation
4. Image recognition
5. Predicting earthquakes.

2.13. What is batch normalization ?

Ans. Batch normalization is a technique for training every deep neural network that standardizes the inputs to a layer for each mini-batch.

2.14. What is Generative Adversarial Network ?

Ans. Generative Adversarial Networks (GANs) are a powerful class of neural networks that are used for unsupervised learning. GANs are made up of a system of two competing neural network models which compete with each other and are able to analyze, capture and copy the variations within a dataset.

2.15. What are the advantages of batch normalization ?

Ans. **Advantages of batch normalization :**

1. Reduces internal covariant shift.
2. Reduces the dependence of gradients on the scale of the parameters or their initial values.
3. Regularizes the model and reduces the need for dropout.

2.16. What are different types of semi-supervised learning algorithm ?

Ans. **Different types of semi-supervised learning algorithm :**

1. Self training
2. Generative models
3. Co-training
4. Graph based methods
5. Semi-supervised Support Vector Machines (S3VM)

2.17. What are the three parts of GANs ?

Ans.

1. Generative
2. Adversarial
3. Networks

2.18. What are the assumptions about data in semi-supervised algorithm ?

Ans.

1. Continuity assumption
2. Cluster assumption
3. Manifold assumption

☺☺☺

3 UNIT

Dimensionality Reduction (2 Marks Questions)

3.1. What do you mean by dimensionality reduction ?

Ans. Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration. Dimensionality reduction methods include wavelet transforms and principal components analysis, which transform or project the original data onto a smaller space.

3.2. What are the limitations of logistic regression?

Ans. **Limitation of logistic regression :**

1. Two-class problems.
2. Unstable with well-separated classes.
3. Unstable with few examples.

3.3. What do you mean by PCA ?

Ans. Principal Component Analysis (PCA) technique adopted for dimensionality reduction. Using the PCA technique, basically a higher dimensional data space can be transformed onto a lower dimensional space. This transformation is also called the Hotelling transform.

3.4. How does LDA work ?

Ans.

1. Firstly, we need to calculate the separabitity between classes which is the distance between the mean of different classes. This is called the between-class variance

$$S_b = \sum_{i=1}^{g} N_i \,(\bar{x}_i - \bar{x})\,(\bar{x}_i - \bar{x})^T$$

2. Secondly, calculate the distance between the mean and sample of each class it is also called the within-class variance

$$S_w = \sum_{i=1}^{g}(N_i - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{N}(x_{i,j} - \bar{x}_i)\,(x_{i,j} - \bar{x}_i)^T$$

3. Finally, construct the lower-dimensional space which maximizes the between-class variance and minimizes the within-class variance *P* is considered as the lower-dimensional space projection also called fisher's criterion.

$$P_{ida} = \arg \lim_{P} \frac{|P^T S_b P|}{|P^T S_w P|}$$

3.5. What are the components of dimensionality reduction ?

Ans. **There are two components of dimensionality reduction :**

1. Feature selection
2. Feature extraction

3.6. What are the various methods used for dimensionality reduction ?

Ans. **The various methods used for dimensionality reduction include :**

1. Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA)
3. Generalized Discriminant Analysis (GDA)

3.7. Explain the process of LDA.

Ans. **Step 1 :** Computing the within-class and between-class scatter matrices.

Step 2 : Computing the eigen vectors and their corresponding eigen values for the scatter matrices.

Step 3 : Sorting the eigen values and selecting the top *k*.

Step 4 : Creating a new matrix that will contain the eigen vectors mapped to the *k* eigen values.

Step 5 : Obtaining new features by taking the dot product of the data and the matrix from step 4.

3.8. How are LDA models represented ?

Ans. The model consists of the statistical properties of our data that has been calculated for each class. The same properties are calculated over the multivariate gaussian in the case of multiple variables. The multivariates are means and covariate matrix. Predictions are made by providing the statistical properties into the LDA equation. The properties are estimated from our data. Finally the model values are saved to file to create the LDA model.

3.9. What are the applications of LDA ?

Ans. **Applications of LDA :**

1. Face recognition
2. Medical
3. Customer identification

3.10. What is undercomplete autoencoder ?

Ans. An autoencoder whose code dimension is less than the input dimension is called undercomplete. Learning an undercomplete representation forces the autoencoder to capture the most salient features of the training data.

3.11. Define regularized autoencoder.

Ans. Regularized autoencoders provide the ability to train any architecture of autoencoder successfully by choosing the code dimension and the capacity of the encoder and decoder based on the complexity of distribution to be modeled. It uses a loss function that encourages the model to have other properties besides the ability to copy its input to its output.

3.12. What is sparse autoencoder ?

Ans. A sparse autoencoder is an autoencoder whose training criterion involves a sparsity penalty $\Omega(h)$ on the code layer h, in addition to the reconstraction error :

$$L(x, g(f(x))) + \Omega(h)$$

where $g(h)$ is the decoder output and typically we have $h = f(x)$, the encoder output.

3.13. Discuss denoising autoencoder.

Ans. Denoising Autoencoder (DAE) is an autoencoder that receives a corrupted data point as input and is trained to predict the original, uncorrupted data point as its output.

3.14. What is PSD ?

Ans. Predictive sparse decomposition (PSD) is a model that is a hybrid of sparse coding and parametric autoencoders. A parametric encoder is trained to predict the output of iterative inference. PSD has been applied to unsupervised feature learning for object recognition in images and video, as well as for audio. The model consists of an encoder $f(x)$ and a decoder $g(h)$ that are both parametric.

3.15. Give applications of autoencoders.

Ans. **Applications of autoencoders :**

1. Dimensionality reduction
2. Image compression
3. Image denoising
4. Feature extraction
5. Image generation
6. Sequence to sequence prediction
7. Recommendation system

3.16. What is an autoencoders ?

Ans. An autoencoder is a neural network that is trained to attempt to copy its input to its output. Internally, it has a hidden layer that describes a code used to represent the input.

3.17. What are the advantage of PCA ?

Ans.

1. Lack of redundancy of data given the orthogonal components.
2. Reduced complexity in images grouping with the use of PCA.

3.18. What are the advantages of LDA ?

Ans.

1. Completely unsupervised
2. Intuitive
3. Built-in classification

3.19. What are disadvantages of LDA ?

Ans.

1. Not Scalable
2. Inefficient to update a model

☺☺☺

4 UNIT

Optimization and Generalization (2 Marks Questions)

4.1. What is optimization in deep learning ?

Ans. Optimization refers to the task of either minimizing some function $f(x)$ by altering x. We usually phrase most optimization problems in terms of minimixing $f(x)$. Maximization may be accomplished via a minimization algorithm by minimizing $f(x)$.

4.2. Define optimization algorithm.

Ans. Optimization algorithms helps us to minimize (or maximize) an objective function (another name for Error function) $E(x)$ which is a mathematical function dependent on the Model's internal learnable parameters which are used in computing the target values(Y) from the set of predictors(X) used in the model.

4.3. What are different types of optimization algorithm ?

Ans. **Different types of optimization algorithm :**

1. First order optimization algorithms
2. Second order optimization algorithms

4.4. Define stochastic optimization.

Ans. Stochastic optimization refers to a collection of methods for minimizing or maximizing an objective function when randomness is present.

4.5. What is spatial transformer network ?

Ans. Spatial transformer networks are a generalization of differentiable attention to any spatial transformation. Spatial transformer networks allow a neural network to learn how to perform spatial transformations on the input image in order to enhance the geometric invariance of the model.

4.6. What are the applications of recurrent neural network ?

Ans. **Applications of recurrent neural networks :**

1. Language modelling and prediction
2. Speech recognition
3. Machine translation
4. Image recognition and characterization

4.7. What are the major parts of LSTM ?

Ans. **Major parts of LSTM :**

1. Forget gate
2. Input gate
3. Output gate

4.8. What are the application of LSTM ?

Ans. **Applications of LSTM includes :**

1. Language modelling
2. Machine translation
3. Image captioning
4. Question answering chatbots

4.9. What is neuroscience ?

Ans. Neuroscience (or neurobiology) is the scientific study of the nervous system. It is a multidisciplinary branch of biology that combines physiology, anatomy, molecular biology, developmental biology, cytology, mathematical modeling, and psychology to understand the fundamental and emergent properties of neurons and neural circuits.

4.10. Define computational neuroscience.

Ans. Computational neuroscience is the field of study in which mathematical tools and theories are used to investigate brain function.

4.11. What are the applications of deep reinforcement learning ?

Ans. **Application of deep reinforcement learning :**

1. AI toolkits
2. Manufacturing
3. Automotive industry
4. Finance
5. Healthcare

4.12. What do you mean by natural language processing ?

Ans. Natural language processing studies the problems inherent in the processing and manipulation of natural language and natural language understanding devoted to making computer understand statements written in human language.

4.13. Classify NLP.

Ans. The field of NLP is divided into subfields :

1. **NLU (Natural Language Understanding)** which investigates methods of allowing the computer to comprehend instructions given in English.
2. **NLG (Natural Language Generation)** strive that computer produce ordinary English language so that people can understand computers more easily.

☺☺☺

5 UNIT

Case Study and Applications (2 Marks Questions)

5.1. What is ImageNet ?

Ans. ImageNet is a large database or dataset of images which are originally labelled with Synsets of the WordNet lexicon tree. It was designed by academics intended for computer vision research.

5.2. Where is ImageNet useful ?

Ans. **ImageNet is useful :**

1. Object recognition
2. Image classification
3. Object localization

5.3. What are the applications facial recognition ?

Ans. **Applications of facial recognition :**

1. At airports
2. Mobile phone makers in products
3. Colleges in the classroom
4. Social media companies on websites
5. Businesses at entrances and restricted areas

5.4. What are the application of NLP ?

Ans. **Application of NLP :**

1. Healthcare
2. Computerized personal assistants and personal virtual assistance
3. Customer service
4. Sentiment analysis

5.5. Define Naive Bayes algorithm.

Ans. The Naive Bayesian Analysis (NBA) is a classification algorithm that is based on the Bayesian Theorem, with the hypothesis on the feature's independence.

$$P(c \mid x) = \frac{P(c \mid x) \times P(c)}{P(x)}$$

5.6. What is computer vision ?

Ans. Computer vision is a field of artificial intelligence that trains computers to interpret and understand the visual world. Using digital images from cameras and videos and deep learning models, machines can accurately identify and classify objects and then react to what they see.

5.7. What are the steps involved in NLP ?

Ans. **Steps involved in NLP are :**

Step 1 : Sentence segmentation
Step 2 : Word tokenization
Step 3 : Predicting parts of speech for each token
Step 4 : Lemmatization
Step 5 : Identifying stop words
Step 6 : Dependency parsing
Step 7 : Finding noun phrases
Step 8 : Named Entity Recognition (NER)

5.8. Define Bioinformatics.

Ans. Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software.

5.9. What are the tools used in Bioinformatics ?

Ans. **Major category of Bioinformatics tools are :**

1. Homology and similarity tools
2. Protein function analysis
3. Structural analysis
4. Sequence analysis

5.10. What are the various branches of Bioinformatics ?

Ans. **Various branches of Bioinformatics are :**

1. Animal Bioinformatics
2. Plant Bioinformatics
   a. Agricultural Bioinformatics
   b. Horticultural Bioinformatics
   c. Medicinal plants Bioinformatics
   d. Forest plant Bioinformatics

5.11. Define face recognition.

Ans. Face recognition is a method of identifying or verifying the identity of an individual using their face. Face recognition systems can be used to identify people in photos, video, or in real-time.

5.12. What are the issues related with face recognition ?

Ans. **Issues related with face recognition :**

1. Security
2. Ownership

3. Safety
4. Mistaken identity
5. Basic freedoms

5.13. Define scene understanding.

Ans. Scene understanding is the process of perceiving, analyzing and elaborating an interpretation of a 3D dynamic scene observed through a network of sensors.

☺☺☺