USN | | | | | | | | | |

Scheme of Evaluation with solutions
Internal Assessment Test 2 – JUNE 2021

Sub:	Big data Analytics					Sub Code:	17CS82	Branch:	ISE		
Date:	19/06/2021	Duration:	90 min's	Max Marks:	50	Sem / Sec:	VIII A,B			OBE	
Answer any FIVE FULL Questions									MARKS	CO	RBT

		MARKS	CO	RBT
1	**Explain the design principles of ANN by constructing a model representation by single and multilayer ANN. Describe steps built in ANN (Artificial Neural Network)** **5+5**	[10M]	CO4	L2

Artificial Neural Networks (ANN) are inspired by the information processing model of the mind/brain. The human brain consists of billions of neurons that link with one another in an intricate pattern. Every neuron receives information from many other neurons, processes it, gets excited or not, and passes its state information to other neurons.

Just like the brain is a multipurpose system, so also the ANNs are very versatile systems. They can be used for many kinds of pattern recognition and prediction. They are also used for classification, regression, clustering, association, and optimization activities. They are used in finance, marketing, manufacturing, operations, information systems applications, and so on.

ANNs are composed of a large number of highly interconnected processing elements (neurons) working in a multi-layered structures that receive inputs, process the inputs, and produce an output. An ANN is designed for a specific application, such as pattern recognition or data classification,
and trained through a learning process. Just like in biological systems, ANNs make adjustments to the synaptic connections with each learning instance.

ANNs are like a black box trained into solving a particular type of problem, and they can develop high predictive powers. Their intermediate synaptic parameter values evolve as the system obtains feedback on its predictions, and thus an ANN learns from more training data

Business Applications of ANN

Neural networks are used most often when the objective function is complex,and where there exists plenty of data, and the model is expected to improve over a period of time. A few sample applications:

1. They are used in stock price prediction where the rules of the game are extremely complicated, and a lot of data needs to be processed very quickly.
2. They are used for character recognition, as in recognizing hand-written text, or damaged or mangled text. They are used in recognizing finger prints. These are complicated patterns and are unique for each person. Layers of neurons can progressively clarify the pattern leading to a remarkably accurate result.
3. They are also used in traditional classification problems, like approving afinancial loan application.

Design Principles of an Artificial Neural Network

1. A neuron is the basic processing unit of the network. The neuron (or processing element) receives inputs from its preceding neurons (or PEs), does some nonlinear weighted computation on the basis of those inputs, transforms the result into its output value, and then passes on the output to the next neuron in the network . X‘s are the inputs, w‘s are the weights for each input, and y is the output.

2. A Neural network is a multi-layered model. There is at least one input neuron, one output neuron, and at least one processing neuron. An ANN with just this basic structure would be a simple, single-stage computational unit. A simple task may be processed by just that one
neuron and the result may be communicated soon. ANNs however, may have multiple layers of processing elements in sequence. There could be many neurons involved in a sequence depending upon the complexity of the predictive action. The layers of PEs could work in sequence, or they could work in parallel

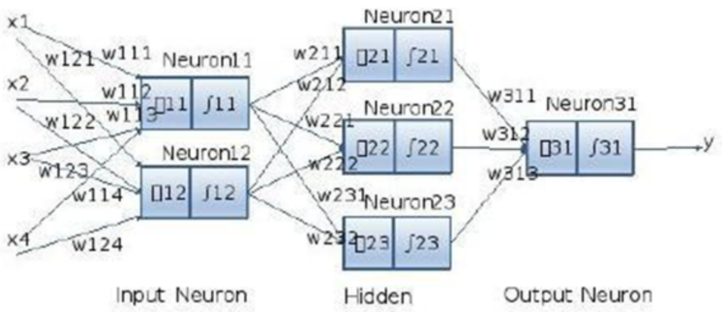


Fig 4.3: Model for a multi-layer ANN

3. The processing logic of each neuron may assign different weights to the various incoming input streams. The processing logic may also use nonlinear transformation, such as a sigmoid function, from the processed values to the output value. This processing logic and the intermediate weight and processing functions are just what works for the system as a whole, in its objective of solving a problem collectively. Thus, neural networks are considered to be an opaque and a black-box system.

4. The neural network can be trained by making similar decisions over and over again with many training cases. It will continue to learn by adjusting its internal computation and communication based on feedback about its previous decisions. Thus, the neural networks become better at making a decision as they handle more and more decisions. Depending upon the nature of the problem and the availability of good training data, at some point the neural network will learn enough and begin to match the predictive accuracy of a human expert. In many practical situations, the predictions of ANN, trained over a long period of time with a large number of training data, have begun to decisively become more accurate than human experts. At that point ANN can begin to be seriously considered for deployment in real situations in real time.

2 **Using Apriori algorithm create the association rules with following data set. Given s= 33% and C = 50%**

Transaction List				
1	**Milk**	**Egg**	**Bread**	**Butter**
2	**Milk**	**Butter**	**Egg**	**Ketchup**
3	**Bread**	**Butter**	**Ketchup**	
4	**Milk**	**Bread**	**Butter**	

[10M] CO5 L3

5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

2> Support level = 33%
Confidence level = 50%

item set's	frequency
milk	9
bread	10
butter	10
egg	3
ketchup	3
Cookies	5

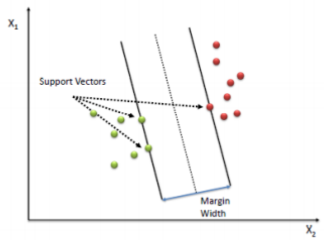
Freq itemsets'	frequency
milk	9
bread	10
Butter	10
Cookies	5

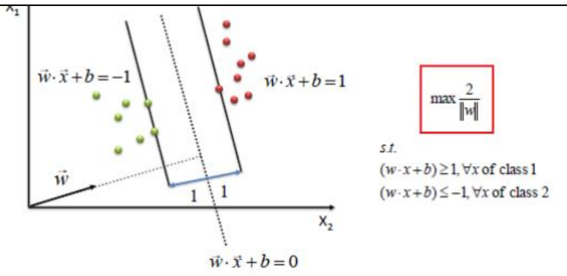
2 item set	freq
milk, bread	7
milk, Butter	7
milk, Cookies	3
bread, butter	9
bread, Cookies	4
Butter, Cookies	3

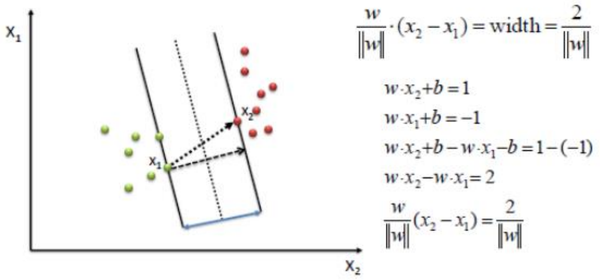
freq 2 items set	freq
Milk, Bread	7
milk, Butter	7
Bread, Butter	9
Bread, Cookies	4

3 items set	freq
milk, bread, butter	6
milk, Butter, cookies	2
Bread, Butter, Cookies	3
milk, bread, cookies	1

freq 3itemset	freq
milk, bread, butter	6

	Support = 6/12 = 50% Confidence = 6/9 = 66.67% Rule 2 : support = 50 % Confidence = 60 % Rule 3 : {butter} = {milk, bread} Support = 50% Confidence = 60% Valid Rule 4 : {milk, bread} = {butter} support = 50 % confidence = 6/7 = 85.7% Rule 5 : {milk, butter} = {bread} Support = 50% Confidence = 85.7 % Valid Rule 6 : {bread, butter} = {Milk} support = 50% confidence = 66.07% Valid			
3 (a)	**What is Support Vector Machine? What are support vectors? Explain Kernel method.** A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.  **Algorithm** 1. Define an optimal hyperplane: maximize margin 2. Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications. Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space. To define an optimal hyperplane we need to maximize the width of the margin (w).	[5M]	CO4	L2

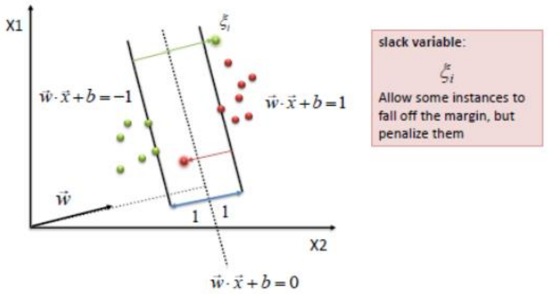




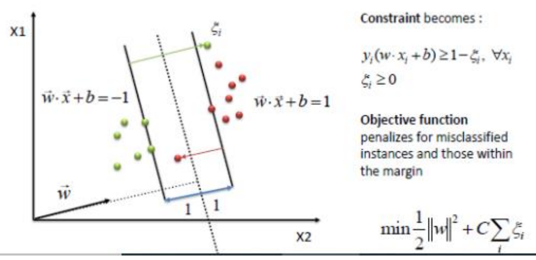
We find w and b by solving the following objective function using Quadratic Programming.

$$\min \frac{1}{2}\|w\|^2$$
$$s.t.\ y_i(w\cdot x_i+b)\geq 1,\ \forall x_i$$

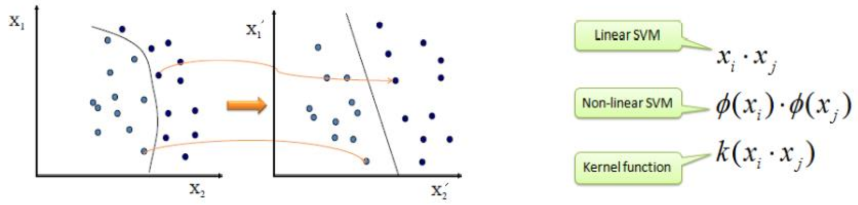
The beauty of SVM is that if the data is linearly separable, there is a unique global minimum value. An ideal SVM analysis should produce a hyperplane that completely separates the vectors (cases) into two non-overlapping classes. However, perfect separation may not be possible, or it may result in a model with so many cases that the model does not classify correctly. In this situation SVM finds the hyperplane that maximizes the margin and minimizes the misclassifications



The algorithm tries to maintain the slack variable to zero while maximizing margin. However, it does not minimize the number of misclassifications (NP-complete problem) but the sum of distances from the margin hyperplanes.



The simplest way to separate two groups of data is with a straight line (1 dimension), flat plane (2 dimensions) or an N-dimensional hyperplane. However, there are situations where a nonlinear region can separate the groups more efficiently. SVM handles this by using a kernel function (nonlinear) to map the data into a different space where a hyperplane (linear) cannot be used to do the separation. It means a non-linear function is learned by a linear learning machine in a high-dimensional feature space while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space. This is called kernel trick which means the kernel function transform the data into a higher dimensional feature space to make it possible to perform the linear separation.



Map data into new space, then take the inner product of the new vectors. The image of the inner product of the data is the inner product of the images of the data. Two kernel functions are shown below.

Polynomial

$$k(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i . \mathrm{x}_j)^d$$

Gaussian Radial Basis function

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{\|\mathrm{x}_i - \mathrm{x}_j\|^2}{2\sigma^2}\right)$$

(b) What is Web Mining? Explain its characteristics and three types of web mining. [5M] CO4 L2

Web mining is the art and science of discovering patterns and insights from the World-wide web so as to improve it. The world-wide web is at the heart of the digital revolution. More data is posted on the web every day than was there on the whole web just 20 years ago. Billions of users are using it every day for a variety of purposes. The web is used for electronic commerce, business communication, and many other applications. Web mining analyzes data from the web and helps find insights that could optimize the web content and improve the user experience. Data for web mining is collected via Web crawlers, web logs, and other means.

Here are some characteristics of optimized websites:

1. Appearance: Aesthetic design. Well-formatted content, easy to scan and navigate. Good color contrasts.
2. Content: Well-planned information architecture with useful content. Fresh content. Search engine optimized. Links to other good sites.
3. Functionality: Accessible to all authorized users. Fast loading times. Usable forms. Mobile enabled. This type of content and its structure is of interest to ensure the web is easy to use. The analysis of web usage provides feedback on the web content, and also the consumer’s browsing habits. This data can be of immense use for commercial advertising, and even for social engineering. The web could be analyzed for its structure as well as content. The usage pattern of web pages could also be analyzed.

Depending upon objectives,
web mining can be divided into three different types:

1. Web usage mining As a user clicks anywhere on a webpage or application, the action is recorded by many entities in many locations. The browser at the client

	machine will record the click, and the web server providing the content would also make a record of the pages served and the user activity on those pages. The entities between the client and the server, such as the router, proxy server, or ad server, too would record that click 2. Web content mining :A website is designed in the form of pages with a distinct URL (universal resource locator). A large website may contain thousands of pages. These pages and their content is managed using specialized software systems called Content Management Systems. Every page can have text, graphics, audio, video, forms, applications, and more kinds of content including user generated content. 3. Web structure mining The Web works through a system of hyperlinks using the hypertext protocol (http). Any page can create a hyperlink to any other page, it can be linked to by another page. The intertwined or self-referral nature of web lends itself to some unique network analytical algorithms. The structure of Web pages could also be analyzed to examine the pattern of hyperlinks among pages. There are two basic strategic models for successful websites: Hubs and Authorities.			
4 (a)	**Explain Naïve Bayes model to classify the data into right class using following data set**	[6M]	CO5	L2

Text	Tag
"A great game"	**Sports**
"The election was over"	**Not sports**
"Very clean match"	**Sports**
"A clean but forgettable game"	**Sports**
"It was a close election"	**Not sports**

Now, which tag does the sentence *A very close game* belong to?

Now we need to transform the probability we want to calculate into something that can be calculated using word frequencies. For this, we will use some basic properties of probabilities, and Bayes' Theorem. If you feel like your knowledge of these topics is a bit rusty, read up on it, and you'll be up to speed in a couple of minutes.

Bayes' Theorem is useful when working with conditional probabilities (like we are doing here), because it provides us with a way to reverse them:

In our case, we have P (Sports | a very close game), so using this theorem we can reverse the conditional probability:

Since for our classifier we're just trying to find out which tag has a bigger probability, we can discard the divisor —which is the same for both tags— and just compare with

This is better, since we could actually calculate these probabilities! Just count how many times the sentence *"A very close game"* appears in the *Sports* tag, divide it by the total, and obtain P (a very close game | Sports).

There's a problem though: "A very close game" doesn't appear in our training data, so this probability is zero. Unless every sentence that we want to classify appears in our training data, the model won't be very useful.

Being Naive

So here comes the *Naive* part: we assume that every word in a sentence is **independent** of the other ones. This means that we're no longer looking at entire sentences, but rather at individual words. So for our purposes, "this was a fun party" is the same as "this party was fun" and "party fun was this".

We write this as:

This assumption is very strong but super useful. It's what makes this model work well with little data or data that may be mislabeled. The next step is just applying this to what we had before

And now, all of these individual words actually show up several times in our training data, and we can calculate them!

Calculating Probabilities

The final step is just to calculate every probability and see which one turns out to be larger.

Calculating a probability is just counting in our training data.

First, we calculate the *a priori* probability of each tag: for a given sentence in our training data, the probability that it is *Sports* P (Sports) is $\frac{3}{5}$. Then, P (Not Sports) is $\frac{2}{5}$. That's easy enough.

Then, calculating P (game | Sports) means counting how many times the word "game" appears in *Sports* texts (2) divided by the total number of words in *sports* (11). Therefore,

However, we run into a problem here: "close" doesn't appear in any *Sports* text! That means that P (close | Sports) = 0. This is rather inconvenient since we are going to be multiplying it with the other probabilities, so we'll end up with

This equals 0, since in a multiplication if one of the terms is zero, the whole calculation is nullified. Doing things this way simply doesn't give us any information at all, so we have to find a way around.

How do we do it? By using something called Laplace smoothing: we add 1 to every count so it's never zero. To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1. In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

Since the number of possible words is 14 (I counted them!), applying smoothing we get tha

The full results are:

Word	P (word | Sports)	P (word | Not Sports)
a	(2 + 1) ÷ (11 + 14)	(1 + 1) ÷ (9 + 14)
very	(1 + 1) ÷ (11 + 14)	(0 + 1) ÷ (9 + 14)
close	(0 + 1) ÷ (11 + 14)	(1 + 1) ÷ (9 + 14)
game	(2 + 1) ÷ (11 + 14)	(0 + 1) ÷ (9 + 14)

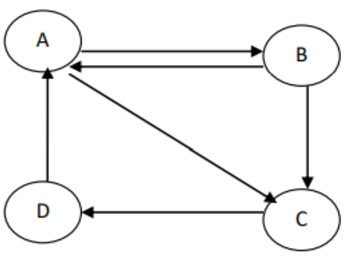
Now we just multiply all the probabilities, and see who is bigger:

Excellent! Our classifier gives "A very close game" the **Sports** tag.

(b) Compare text mining and data mining. [4M] CO5 L2

Below is a table of differences between Data Mining and Text Mining:

S.No.	Data Mining	Text Mining
1.	Data mining is the statistical technique of processing raw data in a structured form.	Text mining is the part of data m which involves processing of tex documents.
2.	Pre-existing databases and spreadsheets are used to gather information.	The text is used to gather high q information.
3.	Processing of data is done directly.	Processing of data is done lingui
4.	Statical techniques are used to evaluate data.	Computational linguistic princip used to evaluate text.
5.	In data mining data is stored in structured format.	In text mining data is stored in unstructured format.
6.	Data is homogeneous and is easy to retrieve.	Data is heterogeneous and is not retrieve.
7.	It supports mining of mixed data.	In text mining, mining of text is done.
8.	It combines artificial intelligence, machine learning and statistics and applies it on data.	It applies pattern recognizing an language processing to unstructu
9.	It is used in fields like marketing, medicine, healthcare.	It is used in fields like bioscienc customer profile analysis.

5	**Compute the Rank values for the nodes for the following network. Which the highest rank node after computation?** 	[10M]	CO5	L4

Solution :

a) Compute the Influence matrix (rank matrix)

- Assign the variables for influence value for each node, as Ra, Rb, Rc, Rd.
- There are two bound links from node A to nodes B and C. Thus, both B and C receives half of node A's influence. Similarly, there are two outbound links from node B to nodes C and A, So both C and A received half of node B's influence.

$$Ra = 0.5*Rb + Rd$$
$$Rb = 0.5*Ra$$
$$Rc = 0.5*Ra + 0.5*Rb$$
$$Rd = Rc$$

	Ra	Rb	Rc	Rd
Ra	0	0.5	0	1.0
Rb	0.5	0	0	0
Rc	0.5	0.5	0	0
Rd	0	0	1.0	0

b) Set the initial set of rank values such as 1/n (n is number of nodes). As 4 nodes are there, initial rank values for all nodes are ¼ i.e 0.25

Variables	Initial Values
Ra	0.25
Rb	0.25
Rc	0.25
Rd	0.25

c) Compute the rank values for 1st iteration and then iteratively compute new rank values till they stabilized.

Variables	Initial Values	Iteration 1
Ra	0.25	0.375
Rb	0.25	0.125
Rc	0.25	0.250
Rd	0.25	0.250

Variables	Initial Values	Iteration 1	Iteration 2
Ra	0.25	0.375	0.3125
Rb	0.25	0.125	0.1875
Rc	0.25	0.250	0.250
Rd	0.25	0.250	0.250

Variables	Initial Values	Iteration 1	Iteration 2	--------	Iteration 8
Ra	0.25	0.375	0.3125	…….	0.333
Rb	0.25	0.125	0.1875	……	0.167
Rc	0.25	0.250	0.250	……	0.250
Rd	0.25	0.250	0.250	…….	0.250

The Final rank shows of node A is highest at 0.333

6 (a)	**What is social network analysis (SNA)? How is it different from other data**	[4M]	CO5	L2

	mining techniques? Social network analysis (SNA) is the process of investigating social structures using networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, information circulation, friendship and acquaintance networks, business networks, social networks, collaboration graphs. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines. These visualizations provide a means of qualitatively assessing networks by varying the visual representation of their nodes and edges to reflect attributes of interest.			
(b)	**Discuss the applications and practical consideration of Social Network Analysis.** Accelerate diffusion by identifying opinion leaders • Reveal how infections spread among patients and staff in a hospital • Map executive's personal network based on email flows • Map interactions amongst blogs on various topics • Map communities of expertise in various fields • Discover emergent communities of interest amongst faculty at various universities • Discover useful patterns in click streams on the WWW • Viral spread: disease, fads and fashions, ideas, YouTube videos • To Find Subject Matter Experts in Particular Area	[6M]	CO5	L2