ANNA UNIVERSITY

CHENNA - 25

Apr - May 2021 Regular Examinations

Rollno : 2018503060

Name : A S SIVA HARI

Program & Branch : BE CSE

Date of Exam : 19/08/21 Session : ~~FN~~/AN ;

Subject code : CS6301 Subject Title : MACHINE LEARNING

PART A			PART B and C						Grand Total
Question number	✓	Marks	Question number	✓	(i) Marks	(ii) Marks	(iii) Marks	Total marks	(in words)
1			11	a					
2				b					
3			12	a					
4				b					
5			13	a					
6				b					
7			14	a					Grand Total:
8				b					
9			15	a					
10				b					
			16						
TOTAL									

DATE : NAME OF THE EXAMINER : SIGN OF EXAMINER :

PART-A

1.

Supervised Learning	Unsupervised learning
(i) Supervised learning algorithms are trained using labeled data	(i) Unsupervised learning algorithms are trained using unlabeled data
(ii) Supervised learning model takes direct feedback to check if it is predicting output or not.	(ii) Unsupervised learning model does not take any feedback
(iii) Supervised learning can be categorized in classification, regression problems	(iii) They can be classified in clustering and association problems

2. Correlation between the decision boundary and the weight vectors (a) If two weight vectors have different norms, the decision boundary is drawn leaning towards the weight vector with smaller norm (b) If they have identical norms, the decision boundary is drawn in the middle.

3 i) Momentum can smooth the progression of the learning algorithm that, in turn can accelerate the training process.

ii) Momentum is set to a value greater than 0.0 and less than one, where common values such as 0.9 and 0.99 are used in practise. Common values of [momentum] used in practise include 0.5, 0.9 and 0.99.

(iii) A validation dataset is a sample of data held back from training the model that is used to give an estimate of model skill tuning model's hyperparameters.

4.

MLP	RBF
(i) The activation function can be any non-linear function which can serve the purpose	(i) The activation function is a function of the euclidean distance of input vector and a certain vector.
(ii) The final layer in MLP also uses the activation function before linearly combining it.	(ii). The final year of RBF don't use activation function. it rather linearly combines the output of the previous neuron.

(ii) There can be more than one hidden layer in MLP	(iii) The final layer have only one neuron.

5. Factor analysis

Factor analysis is one of the unsupervised machine learning algorithms which is used for dimensionality reduction. This algorithm creates factors from the observed variables to represent the common variance i.e variance due to correlation among the observed variables.

6. SVM are applied on binary classification, dividing data points either in 1 or 0. For multiclass classification, the same principle is utilized. The multiclass problem is broken down to multiple binary classification cases, which is also called one-vs-one. At basically divides in class x and rest.

The number of classifiers necessary for one-vs-one multiclass classification can be retrieved with following formula (n being no. of classes)

$$\boxed{\frac{n \times (n-1)}{2}}$$

7. A Markov process is a random process in which the future is independent of the past, given the present. Thus, Markov processes are the natural stochastic analogs of the deterministic processes described by differential and difference equations. They form one of the most important classes of random processes.

---

8. Bagging and Boosting

(i) Bagging and Boosting are ensemble methods focused on getting N learners from a single learner

(ii) Bagging and Boosting make random sampling and generate several training data sets

(iii) Bagging is a method or merging the same type of predictions. Boosting is a method of merging different types of prediction

---

9. Sensitivity is the metric that evaluates a model's ability to predict true positives of each available category.

Specificity is the metric that evaluates a model's ability to predict true negatives of each available category.

$$\text{Sensitivity} : \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

$$\text{Specificity} : \frac{\text{True negatives}}{\text{True negatives} + \text{False positives.}}$$

10. Recurrent neural networks, also known as RNN's are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. RNN's are specialized for processing a sequence of values. They are distinguished by their "memory" as they take information from prior inputs to influence the current input and output.

Types based on input and output

(1) One to one
(2) One to many
(3) Many to many
(4) Many to one.

Applications of RNN model:

1) Language translation
2) Music generation
3) Speech Recognition

11. No of Instances ⇒ no of categories (education qualification)
= 2
⇒ no. of categories (exp)
= 2
⇒ no. of categories (technical)
= 2
⇒ no. of categories (gender)
= 2
⇒ no of categories (marital) = 3

∴ No of instances = $2 \times 2 \times 2 \times 2 \times 3$
= 48.

Find(S) algorithm

1) start with most specific hypothesis
2) If the example is positive, we make the hypothesis and general, when current hypothesis is too specific
3) If example is negative, no of changes are done
4) Repeat it for all example.

Initial hypothesis = (ϕ, ϕ, ϕ, ϕ, ϕ)

↓ 1st training example

(pg, yes, good, male, married)

↓ 2nd example

(pg, yes, good, ?, ?)

↓ 3rd example

No change

↓ 4th example

(pg, yes, good, ?, ?) → Final hypothesis.

Syntactically distinct hypothesis

$= 4 \times 4 \times 4 \times 4 \times 5 = 1280$

Semantically distinct hypothesis

$= (3 \times 3 \times 3 \times 3 \times 4) + 1$

$= 325.$

~~List~~ If-then Elimination Algorithm

1) Create version space of all hypothesis
2) for each training example
   remove hypothesis h(x) for which
   $h(x) \neq c(x)$ where x is training
   example.
3) Output contain the list of hypothesis for dataset.

The algorithm consumes time as it has to generate all possible hypothesis in the versions space. when the data set is small or insufficient the algorithm might output the active set of hypothesis in the version space.

14. RBF network for solving a XOR problem

Gaussian functions :

We need 4 RBF since hidden space is 4D.

$C_{11} \rightarrow t_1 = (0,0), \sigma_1$

$C_{12} \rightarrow t_2 = (0,1), \sigma_2$

$C_{13} \rightarrow t_3 = (1,0), \sigma_3$

$C_{14} \rightarrow t_4 = (1,1), \sigma_4$

$$\sigma_i = \sqrt{\frac{1}{P}\sum_{k=1}^{P}(t_i - t_k)^2}$$

where p = neighbours

$p = 2$ (here) given

$$\sigma_1 = \sqrt{\frac{1}{2}[(t_1 - t_3)^2 + (t_1 - t_2)}$$

$$= \sqrt{\frac{1}{2}\{[(1-0)^2 + (0-1)^2] + [(1-0)^2 + (0-0)^2]}$$

$\boxed{\sigma_1 = 1}$

Similarly,

$\sigma_2 = 1$

$\sigma_3 = 1$

$\sigma_4 = 1$

$\therefore$ $C_{11}(x) = e^{-\frac{|(x - t_1)|^2}{2}}$

$C_{12}(x) = e^{-\frac{|(x - t_2)|^2}{2}}$

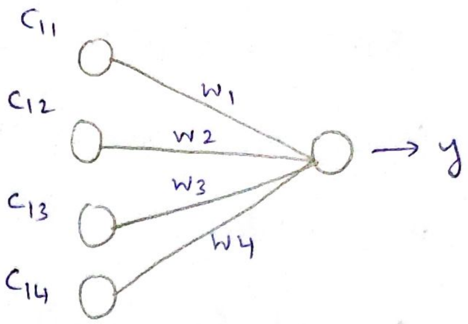
$C_{13}(x) = e^{-\frac{|(x - t_3)|^2}{2}}$

$C_{14}(x) = e^{-\frac{|(x - t_4)|^2}{2}}$

I/P	C_{11}	C_{12}	C_{13}	C_{14}	$\sum w_{kj} c_j$	y
0, 0	1	0.6	0.6	0.4	0.2	1
0, 1	0.6	1	0.4	0.6	-0.2	0
1, 0	0.6	0.4	1	0.6	-0.2	0
1, 1	0.4	0.6	0.4	1	0.2	1

$w_{11} = 1$, $w_{12} = -1$, $w_{13} = -1$, $w_{14} = 1$

$\Sigma w_{kj} O_j$ =



$= (1 \times 1) + (1 \times 0.6)$

$+ (-1 \times 0.6) + (1 \times 0.4)$

$= 1 - 0.6 - 0.6 + 0.4$

$= 0.2$

$= 1 \times 0.6 + (-1) + (0.4 \times -1) + (0.6 \times 1)$

$= 0.6 - 1 - 0.4 + 0.6$

$= -0.2$

$y = C_{1a} (\Sigma W_{kj} O_j)$

$j = 1 \dots 4$

$k = 1$ (1 o/p).

20. Build ID3 decision tree:

Formulae: Entropy $(p) = -\sum_i p_i \log_2 p_i$

$$\text{Gain}(S, F) = \text{Entropy}(S) - \sum_{f \in values(F)} \frac{|S_f|}{|S|} \text{Entropy}(S_f)$$

Sol: Let E be entropy and G be the information gain

$$E(S) = -\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8}$$

$\Rightarrow E(S) = 0.811$

Let us calculate gain for every attribute,

G (S, Educational qualification)

$$= E(S) - \frac{|S_{pg}|}{|S|} E(S_{pg}) - \frac{|S_{ug}|}{|S|} E(S_{ug})$$

$$= 0.811 - \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) - \frac{4}{8}\left(-\frac{4}{4}\log_2\frac{4}{4} - 0/4 \log_2 0/4\right)$$

$= 0.311$

$$G(S, \text{Exp} > 5\text{year}) = E(S) - \frac{|S_{yes}|}{|S|} E(S_{yes}) - \frac{|S_{no}|}{|S|} E(S_{no})$$

$$= 0.811 - \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) - \frac{4}{8}\left(-\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4}\right)$$

$= 0.311$

$$G(S, \text{technical}) = E(S) - \frac{|S_{good}|}{|S|} E(S_{good}) - \frac{|S_{bad}|}{|S|} E(S_{bad})$$

$$= 0.811 - \frac{6}{8}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right) - \frac{2}{8}\left(-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right)$$

$= 0.123.$

$$G(S, \text{gender}) = E(S) - \frac{|S_{female}|}{|S|} E(S_{female}) - \frac{|S_{male}|}{|S|} E(S_{male})$$

$$= 0.811 - \frac{5}{8}\left(-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}\right) - \frac{3}{8}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)$$

$= 0.016.$

$$G(S, \text{marital status}) = E(S) - \frac{|S_{married}|}{|S|} E(S_{married}) - \frac{|S_{single}|}{|S|} E(S_{single}) - \frac{|S_{window}|}{|S|} E(S_{window})$$

$$= 0.811 - \frac{4}{8}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) - \frac{2}{8}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$-\frac{2}{8}\left(-\frac{0}{2}\log_2\frac{0}{2}-\frac{2}{2}\log_2\frac{2}{2}\right)$

$= 0.156.$

As we can see educational qualification and exp > 5 years have highest gain. We can choose any one. Let us choose educational qualification as the root.

Repeat the process for every subtree

$E(S_{pg}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

$G(S, Exp > 5) = 1 - \frac{|S_{yes}|}{S} E(S_{yes}) - \frac{|S_{no}|}{|S|} E(S_{no}) = 1$

$G(S, \text{technical}) = 1 - \frac{3}{4}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) - \frac{1}{4}\left(-\frac{1}{1}\log_2\frac{1}{1}\right)$

$= 0.311$

Similarly,

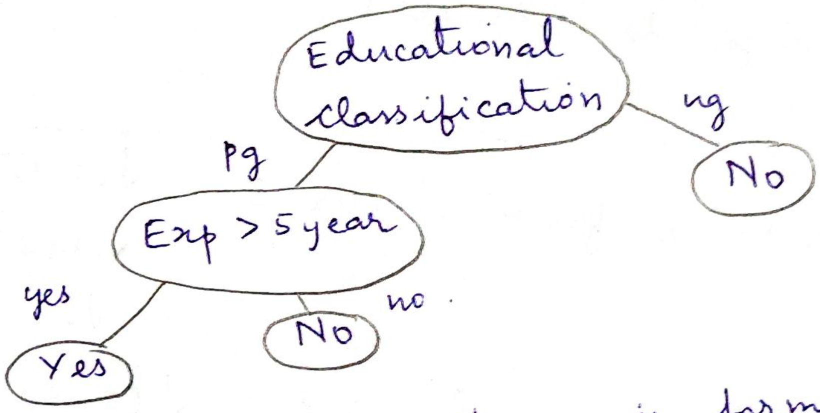
$G(S, \text{marital}) = 0$, $G(S, \text{gender}) = 0$

∴ Exp > 5 years has the highest gain

As we can see, this splits the data into values of some clear

∴ Both can be written as a leaf

∴ The ID3 tree formed is :



Hence, the decision tree is formed for the given data.

---

21. K Means clustering.

labelling the points

A (1,1), B (2,2), C (1,2), D (2,1), E (5,5), F (5,6)

G (6,5), H (6,6)

K = 2

Taking the initial centroids A (1,1) and E (5,5) for clusters 1 and 2 respectively.

Iteration 1

A(1,1) Distance from cluster 1 $= \sqrt{(1-1)^2 + (1-1)^2}$
$= 0$.

Distance from cluster 2 $= \sqrt{(1-5)^2 + (1-5)^2}$
$= \sqrt{16+16}$
$= 4\sqrt{2}$

'A' belongs to cluster 1.

Similarly,

B(2,2) Distance from cluster 1 $= \sqrt{(2-1)^2 + (2-1)^2}$
$= \sqrt{1+1}$
$= \sqrt{2}$

Distance from cluster 2 $= \sqrt{(2-5)^2 + (2-5)^2}$
$= \sqrt{9+9}$
$= 3\sqrt{2}$

'B' belongs to cluster 1.

C(1,2) cluster 1 $= \sqrt{(1-1)^2 + (2-1)^2}$
$= 1$

cluster 2 $= \sqrt{(1-5)^2 + (2-5)^2}$
$= \sqrt{16+9}$
$= 5$.

D(2,1): Cluster 1 : $\sqrt{(2-1)^2 + (1-1)^2}$
$= \sqrt{1}$
$= 1$//

Cluster 2 : $\sqrt{(2-5)^2 + (1-5)^2}$
$= \sqrt{9+16} = 5$.

'D' belongs to cluster 1.

E (5, 5) : cluster 1 = $\sqrt{(5-1)^2 + (5-1)^2}$

$= \sqrt{16+16} = 4\sqrt{2}$

cluster 2 = $\sqrt{(5-5)^2 + (5-5)^2} = 0$.

'E' belongs to cluster 2.

F (5, 6) cluster 1 = $\sqrt{(5-9)^2 + (6-1)^2}$

$= \sqrt{16 + 25} = \sqrt{41}$

cluster 2 = $\sqrt{(5-5)^2 + (6-5)^2} = 1$.

'F' belongs to cluster 2

G (6, 5) cluster 1 : $\sqrt{(6-1)^2 + (5-1)^2}$

$= \sqrt{25+16}$

$= \sqrt{41}$

Cluster 2 : $\sqrt{(6-5)^2 + (5-5)^2}$

$= 1$.

'G' belongs to cluster 2

H (6, 6)

cluster 1 : $\sqrt{(6-1)^2 + (6-1)^2}$

$= \sqrt{25+25}$

$= 5\sqrt{2}$

cluster 2 : $\sqrt{(6-5)^2 + (6-5)^2}$

$= \sqrt{1+1}$

$= \sqrt{2}$

'H' belongs to cluster 2

After iteration 1 :

cluster 1 : A (1, 1), B (2, 2), C (1, 2), D (2, 1)

cluster 2 : E (5, 5), F (5, 6), G (6, 5), H (6, 6)

Ashy

updating Centroids

cluster 1 = $\left(\frac{1+2+1+2}{4}, \frac{1+2+2+1}{4}\right) = (1.5, 1.5)$

cluster 2 = $\left(\frac{5+5+6+6}{4}, \frac{5+6+5+6}{4}\right) = (5.5, 5.5)$

<u>Iteration 2</u> :

A (1, 1) cluster 1 = $\sqrt{(1-1.5)^2 + (1-1.5)^2}$

$= \sqrt{0.25 + 0.25} = \sqrt{0.5}$

cluster 2 = $\sqrt{(1-5.5)^2 + (1-5.5)^2}$

$= \sqrt{20.25 + 20.25} = \sqrt{40.5}$

'A' belongs to cluster 1.

B (2, 2) cluster 1 : $\sqrt{(2-1.5)^2 + (2-1.5)^2} = \sqrt{0.25 + 0.25}$

$= \sqrt{0.5}$

cluster 2 : $\sqrt{(2-5.5)^2 + (2-5.5)^2} = \sqrt{12.25 + 12.25}$

$= \sqrt{24.5}$

'B' belongs to cluster 1.

~~cd(4,2)~~ C (1, 2) cluster 1 $\sqrt{(1-1.5)^2 + (2-1.5)^2}$

$= \sqrt{0.25 + 0.25}$

$= \sqrt{0.5}$

cluster 2 : $\sqrt{(1-5.5)^2 + (2-5.5)^2} = \sqrt{20.25 + 12.25}$

$= \sqrt{32.5}$

'C' belongs to cluster 1.

D (2, 1) cluster 1 : $\sqrt{(2-1.5)^2 + (1-1.5)^2} = \sqrt{0.25 + 0.25}$

$= \sqrt{0.5}$

cluster 2 : $\sqrt{(2-5.5)^2 + (1-5.5)^2} = \sqrt{12.25 + 20.25}$.

$= \sqrt{32.5}$

'D' belongs to cluster 1.

E (5, 5) cluster 1 : $\sqrt{(5-1.5)^2 + (5-1.5)^2} = \sqrt{12.25 + 12.25}$

$= \sqrt{24.5}$

cluster 2 : $\sqrt{(5-5.5)^2 + (5-5.5)^2} = \sqrt{0.25 + 0.25}$

$= \sqrt{0.5}$

'E' belongs to cluster 2.

F (5, 6) : cluster 1 $\sqrt{(5-1.5)^2 + (6-1.5)^2} = \sqrt{12.25 + 20.25}$

$= \sqrt{32.5}$

cluster 2 : $\sqrt{(5-5.5)^2 + (6-5.5)^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5}$

'F' belongs to cluster 2.

G (6, 5) cluster 1 = $\sqrt{(6-5.5)^2 + (5-5.5)^2} = \sqrt{0.25 + 0.25}$

$= \sqrt{0.5}$

cluster 2 = $\sqrt{(6+1.5)^2 + (5-1.5)^2} = \sqrt{20.25 + 12.25}$

$= \sqrt{32.5}$

'G' belongs to cluster 2.

H (6, 6) : cluster 1 : $\sqrt{(6-1.5)^2 + (6-1.5)^2} = \sqrt{20.25 + 20.25}$

$= \sqrt{40.5}$

cluster 2 : $\sqrt{(6-5.5)^2 + (6-5.5)^2} = \sqrt{0.25 + 0.25}$

$= \sqrt{0.5}$

'H' belongs to cluster 2

clusters have not changed after iteration 2, hence the final clusters are

cluster 1 : (1,1), (2,2), (1,2), (2,1)

cluster 2 : (5,5), (5,6), (6,5), (6,6).

22. CNN model

CNNs are a class of deep Neural Networks that can recognize and classify particular features from images and are widely used for analyzing visual images. Their applications range from image and video recognition, image classification, medical image analysis, computer vision and natural processing.

The term 'Convolution' in CNN denotes the mathematical function of convolution which special kind of linear operation wherein two functions are multiplied to produce a third examination.

Basic Architecture

(i) A convolutional tool that separates and identifies the various features of the image

(ii) A fully connected layer that utilizes the output from the convolution process and predicts the class of the image based on the features extracted in previous stages.

Layers

Input → Convolution → Pooling → Fully connected → output

Convolutional layer

This layer is the first layer that is used to extract various features from the input images.

Pooling layer

In most cases, a Convolutional layer is followed by a Pooling layer. The primary aim of this layer is to decrease the size of the convolved feature map to reduce the computational costs.

Fully Connected layer

The Fully connected (FC) layer consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers.

Dropout

Usually, when all the features are connected to the FC layer, it can cause overfitting in the training dataset. Overfitting occurs when a particular model works so well on the training data causing a negative impact.

Activation Functions

Finally, one of the most important parameters of the CNN model is the activation function. They are used to learn and approximate any kind of continuous and complex between variables of the network.

Here W = 227, H = 227, S = 4, padding = 0, 96 filters and stride = 4.

(W = 57, H = 57, Depth = 96)

12. Candidate elimination algorithm incrementally builds the version space given a hypotheses space H and set of E of examples. The example are added one by one, each example possibly shrinks the version space by removing the hypothesis that is consistent with the example.

Algorithm:

step 1 : Load Data Set

step 2 : Initialize General hypoltesis and Specific hypothesis.

step 3 : For each training example.

step 4 : If example is positive example

if attribute_value == hypothesis_value

Do nothing

else :

replace attribute value with '?'

(Basically generalizing it)

step 5 : If example is negative example

make generalize hypothesis more specific.

e.g G_0 = { ?, ?, ?, ?, ? }

S_0 = { ϕ, ϕ, ϕ, ϕ, ϕ }

S_1 = { pg, yes, Good male, Married }

G_1 = { ?, ?, ?, ?, ? }

S_2 = { pg, yes, Good, ?, ? }

G_2 = { ?, ?, ?, ?, ? }

S_3 = { pg, yes, Good, ?, ? }

G_3 = { <pg, ????>, <? Yes ???>, <?? Good ??> }

Ans ⇒ S_4 = { pg, yes, Good, ?, ? }

G_4 = { <pg, ????>, <? Yes ???>, <?? Good ??> }

17. Calculating probabilities for eligibility (assuming 'Male' in Martial status is meant to be "Married")

$p(no) = 3/4$. $P(yes) = 1/4$

Educational qualification:

$P(pg/yes) = 1$, $P(pg/no) = 1/3$, $P(ug/yes) = 0$, $P(ug/no) = 2/3$

Exp > 5 years

$P(yes/yes) = 1$, $P(yes/no) = 1/3$ $P(no/yes) = 0$, $P(no/no) = 2/3$

Technical

$P(good/yes) = 1$; $P(good/no) = 2/3$, $P(bad/yes) = 0$, $P(bad/no) = 1/3$

Gender

$P(female/yes) = 1/2$, $P(female/no) = 2/3$, $P(male/yes) = 1/2$, $P(male/no) = 1/3$.

Martial status

$P(married/yes) = 1/2$, $P(married/no) = 1/2$, $P(single/yes) = 1/2$
$P(single/no) = 1/6$, $P(widow/yes) = 0$, $P(widow/no) = 1/3$

As an example is not given in the question. I am assuming candidate x to be

$X = \langle ug, yes, good, Male, single \rangle$

Eligible (yes):

$P(x/yes) \cdot P(yes) = P(ug/yes) \times P(yes/yes) \times P(good/yes) \times P(male/yes) \times P(single/yes) \times P(yes)$

$= 0 \times 1 \times 1 \times 1/2 \times 1/4$

$= 0$

Not Eligible (no)

$P(X/no) \cdot P(no) = P(ug/no) \times P(yes/no) \times P(good/no) \times P(single/no) \times P(no)$
$\times P(male/no)$

$= 2/3 \times 1/3 \times 2/3 \times 1/3 \times 1/6 \times 3/4$

$= 1/162 = 0.0062 \; (>0)$

$\boxed{no > yes}$ $(P(no) > P(yes))$

X is not eligible.

---

18. Gaussian Mixture Model

~~suppose~~ ~~There are~~ ~~k clusters~~ Gaussian mixture models are a probablistic model for representing normally distributed sub populations within a overall population.

Use
* Feature extraction from speech data
* Object Tracking of multiple objects where no. of mixture components and their means predict objects locations at each frame in a video sequence

$$P(x_i \in C_m) = \frac{\hat{\alpha}_m \, \phi(x_i ; \hat{\mu}_m ; \hat{\Sigma}_m)}{\sum_{k=1}^{n} \hat{\alpha}_m \, \phi(x_i , \hat{\mu}_k ; \Sigma_k)}$$

$P(x_i \in C_m) \rightarrow$ Probability that x_i belongs to C_m

$\phi(x_i ; \mu_m, \Sigma_m) \rightarrow$ Gaussian function with mean μ_m and covarience matrix Σ_m.

$\alpha_m \Rightarrow$ weights with constraints $\sum_{m=1}^{m} \alpha_m = 1$

The problem is how to choose the weights for α_m. The common approach is to aim for maximum likelihood situation using expectation maximisation algorithm.

Initialization :

1) set $\hat{\mu_1}$ and $\hat{\mu_2}$ in randomly chosen values from the dataset

2) select $\hat{\sigma_1} = \hat{\sigma_2} = \sum_{i=1}^{N} (y_i - \bar{y})^2 / N$ ($\bar{y}$ - mean of entire data set)

Repeat Until convergence

1) Expectation step $\hat{r_1} = \dfrac{\hat{\pi}\,\phi(y_i; \hat{\mu_1}, \hat{\sigma_1})}{\hat{\pi}\,\phi(y_i; \hat{\mu_1}, \hat{\sigma_1}) \oplus \phi(y_i, \hat{\mu_2}, \sigma_2)}$

$i = 1 \ldots n$

2) Maximization (1) $\hat{\mu_1} = \dfrac{\sum_{i=1}^{n} (1 - \hat{r_i}) y_i}{\sum_{i=1}^{n} (1 - \hat{r_i})}$

3) Maximization (2) $\hat{\mu_2} = \dfrac{\sum_{i=1}^{n} \hat{r_i} y_i}{\sum_{i=1}^{n} \hat{r_i}}$

4) Maximization (3) $\hat{\sigma_1} = \dfrac{\sum_{i=1}^{n} (1 - \hat{r_i})(y_i - \hat{\mu_1})^2}{\sum_{i=1}^{n} (1 - \hat{r_i})}$

5) Maximization $\hat{\sigma_2} = \dfrac{\sum_{i=1}^{n} \hat{r_i} (y_i - \hat{\mu_2})^2}{\sum_{i=1}^{n} \hat{r_i}}$

6) Maximization (5) $\hat{\pi_j} = \sum_{i=1}^{n} \frac{\hat{r_i}}{N}$

General EM algorithm

Initialization, Guess parameters $\hat{\theta}^{(0)}$. Repeat with convergence

→ (E-step) Computer the expectation $Q(\theta', \hat{\theta}^{(i)})$
$= E(f(\theta'; D') \mid D, \hat{\theta}^{(i)})$

→ (M-step) Eliminate new parameters.

$\hat{\theta}^{(i+)}$ as $\max_\theta Q(\theta', \hat{\theta}^{(i)})$.

Example: In modeling human heights data height is modelled as a normal distribution for each gender with a mean 5'10" → male, 5'5" → female. Given only the height data and not the gender assignments for each data point the the distributions of heights would follow the sum of 2 scaled (different variances and shifted different mean) normal distributions. A model that makes this assumption is Gaussian Mixture Model (GMM) and hence is suitable for this example.

24. ## Auto-associative Neural Network

Auto associative neural networks are ~~us~~ the types of neural networks whose input and output vectors are identical. There are special kinds of neural networks that are used to ~~stimulate~~ simulate and explore the associative process. Association in this architecture comes from the instruction of a set of simple processing elements called units which are connected through weighted connections

It has 5 layers

Input layer ⟶ Mapping layer ⟶ Bottleneck layer ⟶ De-mapping Layer ⟶ Output Layer

Auto associative neural networks can be used in many fields

* Pattern recognition
* Bio-informatics
* Voice recognition
* Signal validation etc.

Activation function

① ReLU

→ Rectified linear activation function
→ Simple to implement and effective at overcoming functions such as sigmoid and tanH

② Sigmoid

→ sigmoid is same as logistic regression and called logistic function
→ It take continuous values as input. Output values in the range 0 ~~and~~ to 1. The bigger the input the closer value to 1.

③ tanH

→ hyperbolic target activation function is very similar to sigmoid.

" All the answers in the answer script has been written by my own hand writing, no body has helped me "

[signature]
19/8/21.