Name: B.Neeraj Kumar Roll No: 18B81A1225

AML ASSIGNMENT

1) What do you mean by well-posed learning problem? Explain with Example.

Well-posed learning:

A computer program is said to learn from Experience E with respect to some class of Tasks T & performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

→ To have a Well-posed learning problem, three features must be identified

i) Learning Task: the thing you want to learn.

ii) Performance measure: must know when you did bad & when you did good. Often called the critic.

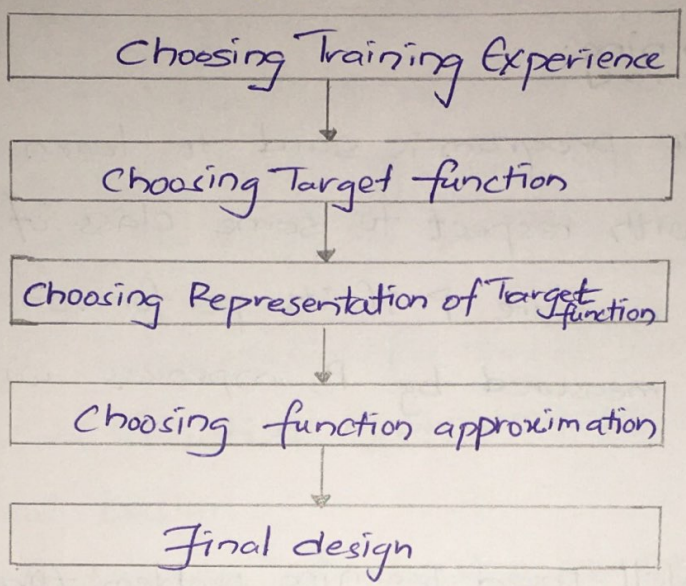
iii) Training Experience: Basically you've got to know how you're going to get the data you're going to train the whole furshluggiver thing with.

Example:

→ A computer program that learns to play checkers might improve its performance as measured by its ability to win at the class of tasks

involving playing checkers game, through experience obtained by playing games against itself.

→ Steps for Designing Learning system are:



1) Choosing Training Experience:

→ The first design choice is to choose the type of training experience from which the system will learn.

→ The type of training experience available can have a significant impact on success or failure of the learner.

→ There are 3 attributes which impact on success or failure of the learner

i) Whether the training experience provides direct or indirect feedback regarding the choices made by performance system.

ii) The degree to which the learner controls the sequence of training examples.

iii) How well it represents the distribution of examples over which the final system performance P must be measured

2) Choosing the Target function:

The next design choice is to determine exactly what type of knowledge will be learned & how this will be used by performance program.

3) Choosing a representation for target function

4) Choosing a function Approximation Algorithm

5) The final Design:

The final design of checkers learning system can be described by four distinct program that represent the central components in many learning systems.

Examples:

→ A checkers learning problem:

- Task T: playing checkers
- Performance measure P: percent of games won against opponents
- Training experience E: playing practice games against itself.

→ A handwriting recognition learning problem:

- Task T: recognizing & classifying handwritten words within images.
- Performance measure P: percent of words correctly classified.
- Training experience E: a database of handwritten words with given classifications

2) Explain the working of Candidate-Elimination algorithm taking enjoy sport concept & training instance given below:

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

→ The candidate-Elimination algorithm computes the version space containing all hypothesis from H that are consistent with an observed sequence of training Examples

→ It begins by initializing the version space to the set of all hypotheses in H; that is by initializing the 'G' boundary set to contain the most general hypothesis in H as

*) $G_0 \leftarrow \{\langle ?, ?, ?, ?, ?, ? \rangle\}$

& initializing the S boundary set to contain the most specific hypothesis as

*) $S_0 \leftarrow \{\langle 0, 0, 0, 0, 0, 0 \rangle\}$

→ General Hypothesis: Not specifying features to learn the machine

→ Specific Hypothesis: Specifying features to learn machine

→ Specific Hypothesis will be same as Find-S algorithm

Algorithm:

1) Initialize General Hypothesis & specific Hypothesis

2) For each training example d, do

→ if d is a positive example.

for each attribute value of specific Hypothesis do

if attribute_value == hypothesis_value:

Do nothing

else:

replace attribute value with '?'

If specific hypothesis is changed then modify general hypothesis.

→ if d is a negative example

Make generalize hypothesis more specific.

Algorithm on given example:

- S0 : (ϕ, ϕ, ϕ, ϕ, ϕ, ϕ) Most specific Boundary
- G0 : (?, ?, ?, ?, ?, ?) Most generic Boundary

→ The first example is a positive example, the hypothesis at specific boundary is inconsistent, hence we extend the specific boundary, & the hypothesis at generic boundary is inconsistent, hence we retain it.

- S1 : (Sunny, Warm, Normal, Strong, Warm, Same)
- G1 : (?, ?, ?, ?, ?, ?)

→ The second example is positive, again the Hypothesis at specific boundary is inconsistent, hence we extend specific boundary, & hypothesis at generic boundary is consistent, hence we retain it.

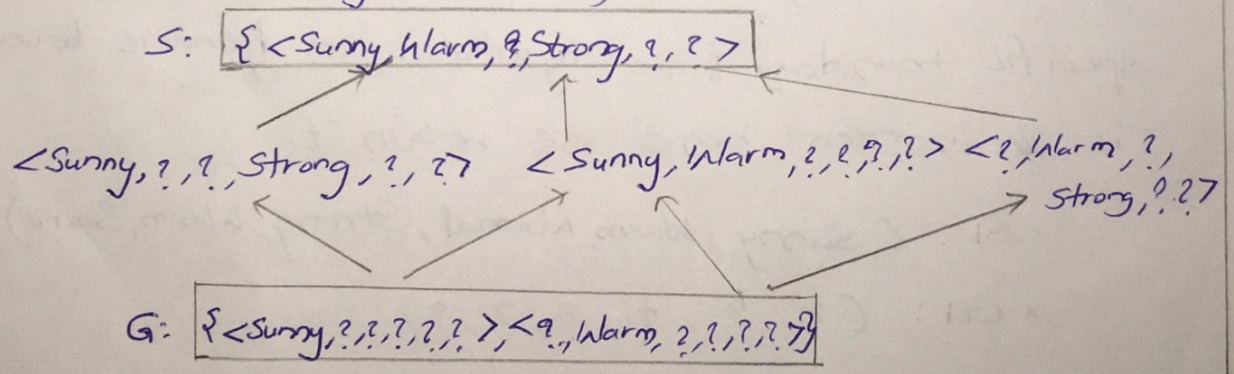
- S2: (Sunny, Warm, ?, Strong, Warm, Same)
- G2: (?, ?, ?, ?, ?, ?)

→ The third example is negative, the hypothesis at specific boundary is consistent, hence we retain it, & hypothesis at generic boundary is inconsistent hence write all consistent hypotheses by removing one "?" at each time.

- S3: (Sunny, Warm, ?, Strong, Warm, Same)
- G3: (Sunny, ?, ?, ?, ?, ?), (?, Warm, ?, ?, ?, ?)
  (?, ?, ?, ?, ?, Same)

→ The fourth example is positive, the hypothesis at specific boundary is inconsistent, hence we extend the specific boundary, & consistent hypothesis at generic boundary is retained

- S4: (Sunny, Warm, ?, Strong, ?, ?)
- G4: (Sunny, ?, ?, ?, ?, ?) (?, Warm, ?, ?, ?, ?)



3) Define Decision Tree learning. Explain decision tree learning algorithm with an example.

Decision Tree learning:

→ Decision Tree learning is one of the most widely used & practical methods for inductive inference

*) Inductive inference is the process of reaching a general conclusion from specific examples.

→ Decision Tree learning is a method for approximating discrete-valued target function.

→ It starts with a root node & ends with a decision made by leaves.

→ To make decision tree we need to Know entropy, information ~~node~~ gain

*) Entropy is nothing but the uncertainity in our dataset.

$$\text{Entropy}(S) = -P_{\oplus}\log_2 P_{\oplus} - P_{\ominus}\log_2 P_{\ominus}$$

*) Information gain measures the reduction of uncertainity

$$\text{Gain} = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{S} \text{Entropy}(S_v)$$

where Values(A) is the all possible values for attribute A & S_v is the subset of S for which attribute A has value v.

	Are ears pointy	Does it bark	label
0	Yes	No	Cat
1	No	Yes	Dog
2	No	Yes	Dog
3	No	Yes	Dog
4	Yes	No	Cat
5	Yes	Yes	Dog
6	No	Yes	Dog
7	Yes	Yes	Dog
8	Yes	No	Cat
9	No	No	Cat

First we take an attribute: Are ears pointy

We should calculate Information gain for different values i.e, Yes or No, lets take Dog as Yes & Cat as No

For finding Information gain we need to calculate Entropy

For attribute : Are years pointy

$S = [5+, 5+]$ $\quad Entropy(S) = \frac{-5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right)$

$= -1 * \log_2(2^{-1}) = 1.0$

$S_{Yes} = [2+, 3-]$ $\quad Entropy(S_{Yes}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.9709$

$S_{No} = [5+, 0-]$ $\quad Entropy(S_{No}) = -\frac{5}{5}\log\left(\frac{5}{5}\right) = 0$

$$Gain(S, \text{Are ears pointy}) = Entropy(S) - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

Gain(S, Are ears pointy) = $1.0 - \frac{5}{10}$ Entropy(S_{Yes}) $- \frac{5}{10}$ Entropy(S_{No})

$= 1.0 - \frac{5}{10}(0.9709) = 0.51455$

For attribute : Does it bark

$S = [6+, 4-]$ Entropy(S) $= -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$

$= 0.9709$

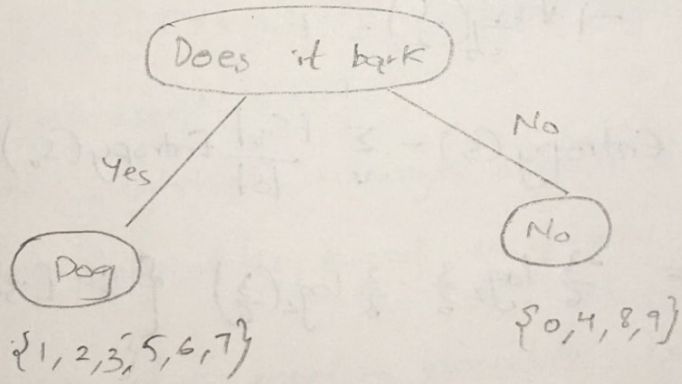
$S_{Yes} = [6+, 0-]$ Entropy(S_{Yes}) $= 0$

$S_{No} = [0+, 4-]$ Entropy(S_{No}) $= 0$

Gain(S, Does it bark) $= 0.9709 - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|}$ Entropy(S_v)

$= 0.9709 - \frac{6}{10}(0) - \frac{4}{10}(0) = 0.9709$

→ The information gain value which has maximum is choosen as root Node



→ If we have multi-valued attribute, then if we can determine them it is directly derives leaf-node otherwise based on the value again the data is further divided to determine the perfect result.

4) Consider the following set of training examples

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T
3	−	T	F
4	+	F	F
5	−	F	T
6	−	F	T

a) What is the entropy of this collection of training examples with respect to the target function classification?

b) What is the information gain of a_2 relative to these training examples?

a) $S = [3+, 3-]$ $\quad Entropy(S) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

$Entropy(S) = -1 * \log_2\left(\frac{1}{2}\right) = 1$

b) $Gain(S, a_2) = Entropy(S) - \sum_v \frac{|S_v|}{|S|} Entropy(S_v)$

$Entropy(S) = \frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$ $\quad \{S = [3+, 3-]\}$

~~Entropy(~~

$S_1 = [2+, 2-] \Rightarrow Entropy(S_1) = \frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$

$S_F = [1+, 1-] \Rightarrow Entropy(S_F) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$Gain(S, a_2) = 1 - \frac{4}{6}(1) - \frac{2}{6}(1) = 0.0$

5) Explain the basic definitions of Sampling theory.

Definitions:

→ A random variable can be viewed as name of an experiment with a probablistic outcome. Its value is the outcome of experiment.

→ A probability distribution for a random variable Y specifies the probability $P_r(Y = y_i)$ that Y will take on the value y_i, for each possible value y_i.

*) Statistical function that describes all the possible values & likelihoods that a random variable can take within a given range.

→ The expected value or mean, of a random variable Y is $E[Y] = \Sigma_k y_i P_r(Y = y_i)$. The symbol μ_Y is commonly used to represent $E[Y]$.

*) The weighted average of all outcomes of that random variables based on their probabilities.

→ The variance of a random variable is $Var(Y) = E[(Y - \mu_Y)^2]$. The variance characterizes the width or dispersion of the distribution about its mean.

→ The standard deviation of Y is $\sqrt{Var(Y)}$. The symbol σ_Y is often used to represent the standard deviation of Y.

→ The Binomial distribution gives the probability of observing r heads in a series of n independent coin tosses. if the probability of heads in a single toss is p.

→ The Normal distribution ~~(can be thought)~~ is a bell-shaped probability distribution that covers many natural phenomena.

→ The Central limit Theorem is a theorem of stating that the sum of large number of independent identically distributed random variables approximately follows a Normal distribution.

→ An estimator is a random variable Y used to estimate some parameter p of an underlying population.

*) An estimator is a statistic that estimates some fact about the population.

→ The estimation bias of Y as an estimator for p is the quantity (E[Y] - p). An unbiased estimator is one for which the bias is zero.

→ A N% confidence interval estimate for parameter p is an interval that includes p with probability N%.

6) Write short notes on the following

a) Binomial Distribution

b) Estimating Hypothesis accuracy.

a) The Binomial distribution gives the probability of observing r heads in a series of n independent toss coins, if the probability of heads in a single toss is p.

→ A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.

→ A binomial distribution is defined by the probability function

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

→ If a random variable X follows a binomial distribution, then:

*) Probability $P_r(X=r)$ that X will take on the value of r is given by $P(r)$

*) the expected, or mean value of X is $E[X] = np$

*) The variance of X, $Var(X) = npq = np(1-p)$

*) The standard deviation of X i.e σ_X is $\sqrt{npq}$

$= \sqrt{np(1-p)}$

b)

→ When evaluating a learning hypothesis involves estimating the accuracy with which it will classify further instances.

→ There are some space of instances X over which target function f may be defined

→ Different instances will have frequencies

→ Based on the frequency, each instance in X will have some unknown probability.

→ Trainer will teach the machine about the training examples of target function.

→ Send a particular instance x_i along with target function $f(x)$

- Target function $f: x \rightarrow \{0,1\}$

→ Classifies each instance into different categories based on requirement.

→ Now after classification what is the probable error in the estimation we made error of 2 types:

- Sample error
- True error.

- Sample Error:
  - → The error rate of hypothesis over a sample of data
  - → The sample error (denoted as $error_S(h)$) of hypothesis h with respect to target function f & data sample S is

$$error_S(h) = \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$

  - → Where n is number of examples in S, & the quantity $\delta(f(x), h(x))$ is 1 if $f(x) \neq h(x)$, and 0 otherwise

- True Error:
  - → The true error of a hypothesis is the probability that it will misclarify a single randomly drawn instance from the distribution D.
  - → The true error (denoted as $error_D(h)$) of hypothesis h with respect to target function f & distribution D, is the probability that h will misclarify an instance drawn at random according to D.

$$error_D(h) = \Pr_{x \in D} [f(x) \neq h(x)]$$