Name: Purushottam Tiwari
Roll No.: 19074029
Branch: CSE(IDD)
Subject: CSE-241(Artificial Intelligence)

Mid-Term Assignment

Ans 1(a). Gradient descent is an iterative process to slowly minimize the cost of our predictions. In linear regression we generally start at a particular value of parameters θ_s and compute predictions for all input examples. We then have the task of calculating the cost of prediction by calculating loss for training sample using the formula for loss:

$$\mathbf{L_i = h(\theta_o) - Y_i}$$

where,
L_i is loss for i^{th} training sample.
$h(\theta_o)$ is prediction with particular values of parameters θ_o.
And, Y_i is the labelling for i^{th} sample in the traning sample.

So cost is given by: $\mathbf{J(\theta_o) = \Sigma(L_i^2)/m}$ which is summation of squared loss for each training sample prediction and m is size of training sample.

The main objective of gradient descent is to find best set of parameter θ_s so that the cost defined above is minimized.
It does so by calculating next set of parameter θ_1 from present set of parameter vector θ_o with the help of differentiation. It uses differeniation to find the direction of maximum change with respect to each Θ in θ_o and then simultaneous updates each Θ in θ_o towards opposite direction of maximum increase to get next set of parameter vector θ_1.
Mathematically, it is given as:

$$\mathbf{\theta_1 = \theta_o - \alpha * J(\theta_o)}$$

where,
α is learning paramter(which determines how far we want to move in our chisen direction).
$J(\theta_o)$ is the cost function calculated at θ_o.
and θ_1 is our required next set of parameters.
Pseudo code:
y’ = theta * X + b

cost = sum([$data^2$ for data in (y-y’)]) / m
m_grad = -(2/N) * sum(X * (y-y’))
b_grad = -(2/N) * sum(y-y’)
theta = theta – (learning_rate*m_grad)
b= b – (learning_rate*b_grad)
...
and then we repeat above process each time updating our parameter vector θ.

Ans 1(b). The intution behind using gradient descent for logistic regression is same as for linear regression. We want to minimize our cost function. The only difference arises in the method using which we calculate our cost function using gradient descent. We use a cross entropy loss function to calculate loss for each sample and then further sum them up to get the total cost of our prediction. So cost function is given by:

$$\mathbf{-1/m*\Sigma[y_i \log\sigma(\theta\cdot x_i+b)+(1-y_i)\log(1-\sigma(\theta\cdot x_i+b))]}$$

The parameter vector θ is then updated using gradient descent to find next set of parameters θ.
Pseudo code for the update of parameter of vector theta is:

theta = theta -learning_rate*(gradient_vector)

Ans 1(c). Various metrics used to evaluate the performance of a amchine learning algorithm are:
(i). Classification Accuracy:
This is the most commonly used metric used for evaluating a machine learning algorithm. It is most of the times simply reffered to as accuracy. It is the ratio of number of correct predictions to the total numer of input samples.
Formula:

accuracy = (Number of correct preditions)/(Total number of predictions made)

It is suitable if our data is fairly distributed for both right and wrong samples. For example if we have a data that contains 60% correctly labelled and 40% wrongly labelled samples. Then this would be a good measure to evaluate our algorithm.
It is not suitable if our data is highly biased towards a particular type.

For example: if we have a data the has 98% labels of class A and 1% labels of class B then we can get 98% accuracy simply by predicting A for every sample. In this case this would not be a very good method to evaluate our algorithm.

(ii). Recall:

Recall is the ratio of the correct positive predictions to total actually positive predictions. More formally it is the ratio of True positives to True positives plus false negatives.

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

Recall is generally used in situations where occurence of false negatives is strictly unwanted. For example: in case of disease(lets say X) prediction we don’t want any person having X to be predicted as not having X as that can be fatal.

(iii). Precision:

It is the ratio of data which is correctly labeled as Positive to all the positive labelled data of our algorithm. More formally it is the ratio of True positives to True positives plus false positives.

$$\frac{\text{True positives}}{\text{True positives + False positives}}$$

It is used in cases where it is desired that we want high value of our true positives and a negligble to zero false positives. For Example: For a spam classifier it is already desired that we do not classify any email which is not spam to be spam as that can cause serious issues with customer satisfaction. So Precision can be used for evaluation in such cases.

(iv). Specificity:

It is the ratio of data that is correctly labelled as negative by our classifier to all the actual negative data. More formally it ti the ration of True negatives to true negatives plus false positives.

$$\frac{\text{True negatives}}{\text{True negatives + False Positives}}$$

It can be used in cases where any false positive is highly undesirable. That is we do not want any data to be wrongly predicted as positives. For

Example: If we are testing drug consumers it is highly undesirable to predict an innocent person as consumer of drug. In such cases specifity provides a good measure to evaluate our algorithm.

(v). F1 Score:

It is harmonic mean of precison and recall. It's range is [0, 1]. It is calculated as:

$$F1 = \frac{2*precision*recall}{precision + recall}$$

where,

precision is ratio of True Positives with total positives predicted by classifier.

recall is the ratio of True positives with all actual positive samples.

It's best for evaluation in cases where we have highly unevenly distributed data.

Higher the value of F1 score better is the performance of classifier.

Ans: 2(a).

For any sentence **α** that has ground term **g,** and for any variable v not occuring in **α** we have,

$$\frac{\alpha}{\exists V Subs_1(\{g/v\}, \boldsymbol{\alpha})}$$

Where $Subs_1$ is a function that substitutes a single occuence of g with v.

Reads (Mary, Hard Times) - ∃x Reads(x, Hard Times)

∀x Reads(x, Hard Times) is equivalent to ¬∃x¬ Reads(x, Hard Times)

Now, ∀ is really a conjunction over the universe of objects and ∃ is a disjuction. So from DeMorgan's rules for quantified and unquantified sentences:

$\forall x \neg P \equiv \neg \exists x P$

$\neg \forall x P \equiv \exists x \neg P$

$\forall x P \equiv \neg \forall x \neg P$

A universally quantified sentence can be replaced by set of all possible instantiations.

So we can write general inference rule, as

$$\frac{P(k) \text{ for some } k}{\exists x\ P(x)}$$

Ans: 2(b).

a. {x/A, y/B, z/B}
Progressive unification:
P(A, B, B), P(x, y, z) : {x/A},

P(A, B, B), P(A, y, z) : {x/A, y/B},

P(A, B, B), P(A, B, z) : {x/A, y/B, z/B},

b. Cannot unify
Progressive unification:
Writes(Story(x), x), Writes(y, y) : {y/Story(x)}

Writes(Story(x), x), Writes(Story(x), Story(x)) : {x/Story(y)} Cannot unify variable x with Story(x) as it is a term referring to variable x. Therefore cannot unify.

Ans: 3(a). It is not a correct proof as the first statement "sun rises in the east" is a universal truth and it has nothing to do with who is evil or good. Character of humans cannot decide how nature conducts it's activities.

Ans: 3(b). It is not a correct proof. As "Sun rises in east and sets in west" these are universal truth and cannot by greatness or evilness of a human.

Ans: 3(c)(i). "Falsehood implies anything", This means if a false statement imples a true statement or a false statement implies false statement i.e. F->T or F-> F then both cases result is True.

Now p->q ≡ ~pvq

as p is false so ~p is T.

Therefore TvT or TvF are always True. So this falsehood implies anything.

Example of a fallacy that can be derived from statement is:

"2 + 2 is 5 so china implies pakistan are same."

As 2 + 2 = 4 so 5 = 4 subtracting 3 both sides.

2 = 1.

Now china and pakistan are two so they are equal to one.

So we see that a fallacy(2+2=5) implied even a false statement(china and pakistan are one).

Ans: 3(c)(ii). If an argument doesn’t proove a conclusino then does not mean that conclusion is wrong.

For the above statement an argument consider an example:

A person is accused of a crime that he has not actually done. But he/she has no prrof to proove that he/she is innocent. So in this It is clearly seen the fallacy that the person is guilty is primarily due to the fact that there are no proofs for him/her to be not guilty because of which it cannot be said that the person is guilty.

Ans: 3(c)(iii). The idea of **validity** of a deductive argument is the most natural approach for formal nature of logic.

And for world knowledge or common sense, the most probable strict property will be **soundness** of an argument.

Proof(lack of soundness of an argument can lead to fallacy): Soundness can be defined as how strong and deep understanding of common this one has. Logically an argument is sound iff:

1. it is valid

2. it is actually true(all premises are true)

Now for knowing that argument is actually true, one should have proper knowledge about the argument otherwise there are chances of raising an error thus declaring the arugment false.

For Example: “I eat food from restaturant, now I am sick so the food in restaurant is unhygenic.”

In this statement the person blames restaurant food for being sick without knowing if the food is really unhygenic. So we can see lack of soundness in statement “restaurant food is unhygenic”(not 100% sure if that is the case) leads to fallacy in the main argument.

So the above example prooves how lack of soundness can eventually lead to fallacy.

General conclusion: We conclude that an statement should be absolutely true or false. For this one requires sound knowledge of common things. Also lack soundness in arguments can eventually lead to fallacy.

Ans: 4(a).

Ontology: These are formal complex definitions of vocabularies which allow us to define complex structures and new relationships between vocabulary terms and members of class and different members of classes that we define. It is a way of showing the properties of a subject area and their relationships, by defining a set of concepts and categories that represent the subject.

Frames: Frames are the AI data structure which divides knowledge into substructure. It consists of a collection of slots and slot values. Slots have names and values which are called facets. These slots can be of any type and sizes. Frames system consist of a collection of frames which are connected.

For xample if we consider a this statement: Banana is a yellow colored fruit.

It's frame representation would be:

Slot	FILTER
Entity	Banana
Color	Yellow
Category	Fruit

In most general cases frame graphs are self sufficient. Facets are features of frames which enable us to put constraints on the frames. When we need to extract information from the frame we can simply search for the particular facet. Therefore there is no need to create an ontology and define a separate relation of that property with frame title.

For example consider information of employee XYZ:

SLOT	FILTER

Name	XYZ
Employee Id	4227
Post	Manager
Address	ABC
Mob No.:	1424285463

Now if we want to search for Post of XYZ, it's simple enough to search for facet "Post" in slots. There is no need of ontology.

Ans 4(b). Categorization is a very good and efficient approach to separate things to study them or understanding them. But in the field of philosophy this categorization becomes a matter of contention. This because of following reasons:

1. In order to understand philosophy in simple then we can call it approach of understanding world and how it should be. But the main problem lies in the fact that every one have there own ideas and aspects of looking in to the world and understanding it.

2. The rationalist tradition objectifies the world in a dramatic way, seeking to model categorizationa s model contention. They have a good influen over AI.

3. Aristotle gave a model of categorisation based on 'neccesary and sufficient conditions'. But here the argument arise that there are a lot of thing which can satisfy more than one conditions or not satisfy any conditions, how to categorise them?

Implications of data driven AI:

1. It focusses on building a system to identify what is the correct answer based on having seen a large amount of pre existing data,

2. It does not utilize set of predefined rules/models and is mainly dependent on huge mount of data.

3. It does not depend on human accurately describing set of rules to every problem.