Muhammad Umar Salman
21010241

Assignment 1

AI 701
September 18, 2021

AI and Intelligent Agents

1. (a) **Intelligence**: Is the quality that an entity possesses so that it can perceive and reason on its own as well as behave and be self-aware according to its environment

   (b) **Artificial Intelligence**:Artificial Intelligence is the intelligence given to an entity so that it can perceive and reason on its own as well as behave and be self-aware according to its environment and operate autonomously while adapting to any change.

   (c) **Agent**: An agent is entity that perceives its environment through sensors and can act on its environment through actuators

   (d) **Rationality**: Rationality is to act rationally. That means based on the performance measure, prior knowledge and of the environment act in the best way/ achieve the best outcome during a moment of uncertainty with whatever limited actions are available.

   (e) **Logical Reasoning**: Logical Reasoning is a way of thinking that uses deductive reasoning given a knowledge base or facts to deduce other facts by reason.

2. Vision Systems are intelligent agents designed by humans to think rationally and logically as well as act in this way. But the matter of the fact is not all real-world problems can be solved on the basis of rationality and logic, humans also think instinctively. For e.g., if a person begins to think rationally or logically to remove his hand from a hot place his hand will burn by the time taken to process all the facts. So just like a human would instinctively remove his hand a human instinctively processes an image without any complex maths. This processing of images or Human "Vision Systems" can in the same way also be thought of as human instinct or human reflex.

3.) One of the first major successes of AI was in 1997 when IBM's Deep Blue was the first computer program to have beat the world chess champion under regular time controls. Gary Kasparov, who had been beaten later said he felt a new type of intelligence. STANLEY, a driverless robotic car, in the 2005 DARPA Grand Challenge completed a 132-mile course at 22mph through the rough terrain of the Mojave Desert winning first prize. The following year in 2006 CMU's BOSS won the Urban Challenge, safely driving in traffic through the streets of a closed Air Force base, obeying traffic rules and avoiding pedestrians and other vehicles. In Speech Recognition a traveler calling United Airlines to book a flight can have the entire conversation guided by an automated speech recognition and dialog management system. Launched in 1996, the Robocup Competition aims to develop a fully autonomous humanoid robots that can win against human world champion soccer team by 2050.
   From these 5 challenges and contests we can see that how it started from winning a chess game against a in 1997 to autonomous cars, fully conversational dialogue system

which answers question to attempting to beat the world soccer champions team in a physical soccer match. From the DARPA Grand Challenge in 2005 in a empty terrain to 2006's Urban Challenge, we can see that how much AI has progressed in the span of only a year where from empty terrain an autonomous car is driving in an urban environment with other traffic. We can also see to what degree these contests have advanced in the state-of-the-art AI when we look at the Robocup's progress from 1996 to 2009 where 43 teams are participating in this contest and they look to robots playing against human world soccer champions by 2050. However, these contests do draw away energy from new ideas as every year each research team is trying to improve or better the AI and AI agent than last year's winner of a certain contest rather than being innovative and creating something new that could maybe solve other problems than just working on the same thing every year.

4. (a) **Agent**: An agent is entity that perceives its environment through sensors and can act on its environment through actuators

   (b) **Agent Function**: An agent function maps a sequence of percept vectors to an action out of a set of finite actions

   (c) **Agent Program**: An Agent Program is a program that runs or implements the agent function on a physical architecture (hardware + software). This program is mostly run on an artificial agent.

   (d) **Rationality**: Rationality is to act rationally. That means based on the performance measure, prior knowledge and perceptions of the environment act in the best way/ achieve the best outcome during a moment of uncertainty with whatever limited actions are available.

   (e) **Autonomy**: Autonomy is the capability of taking independent actions and decision-making. That means an artificial agent should learn from what it can from incorrect or partially incorrect prior knowledge rather than being fully dependent on that prior knowledge only.

   (f) **Reflex Agent**: Reflex Agent is an agent that selects an action based on the current percept, and ignores the rest of the percept history. They are stateless agents and have no memory of past world states

5. (a) **Medical Diagnosis System**

      i. Performance Measure
         A. Maximize Healthiness of Patient
         B. Minimize Cost
         C. Minimize Side Effects of Treatment

      ii. Environment
         A. Patient
         B. Hospital Staff

C. Diagnosis Room

iii. Sensors

A. Touch screen/Keyboard

B. Enter Patient Answers

C. Enter Patient Symptoms

iv. Actuators

A. Ask Diagnosis Questions

B. Display Diagnosis

C. Display Treatments

D. Display tests to be taken

(b) **Satellite Image Analysis System**

i. Performance Measure

A. Maximize Classification Accuracy

B. Minimize computation time

C. Maximize Image Quality

ii. Environment

A. Outer Space

B. Satellite

C. Downlink from satellite

iii. Sensors

A. Camera

B. Color pixel array

iv. Actuators

A. Image Capturer (Takes Images)

B. Display Classification

(c) **Interactive English Tutor**

i. Performance Measure

A. Improve Students English

B. Maximize students score

C. Minimize English mistakes

ii. Environment

A. Students

B. Mock Tests

C. Testing Agency

iii. Sensors

A. Keyboard/Touch Screen

   B. Mic for Vocal Answers
   C. Student Answers
iv. Actuators
   A. Display Corrections
   B. Display Mock Questions
   C. Display Suggestions

Calculus Recap

1. $3x^2 + 1$
2. $f(\theta) = \frac{1}{2}\sum_{i=1}^{n} w_i(\theta - x_i)^2$

   $\frac{d(f(\theta))}{d\theta} = \frac{1}{2}\sum_{i=1}^{n} 2w_i(\theta - x_i)^{2-1} \cdot 1$

   $\frac{d(f(\theta))}{d\theta} = \sum_{i=1}^{n} w_i(\theta - x_i)$

   To minimize θ : $\frac{d(f(\theta))}{d\theta} = 0$ $\sum_{i=1}^{n} w_i(\theta - x_i) = 0$

   $\sum_{i=1}^{n}(\theta w_i - x_i w_i) = 0$

   $\sum_{i=1}^{n} \theta w_i - \sum_{i=1}^{n} x_i w_i = 0$

   $\sum_{i=1}^{n} \theta w_i = \sum_{i=1}^{n} x_i w_i$

   $\theta \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} x_i w_i$

   $\theta = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$

3. $f(w) = \sum_{i=1}^{n}\sum_{j=1}^{n}(a_i^T w - b_i^T w)^2 + \lambda \parallel w \parallel_2^2$

   $f(w) =$
   $(a_1^T w - b_1^T w)^2 + (a_1^T w - b_2^T w)^2 + ... + (a_1^T w - b_n^T w)^2$
   $(a_2^T w - b_1^T w)^2 + (a_2^T w - b_2^T w)^2 + ... + (a_2^T w - b_n^T w)^2$
   :
   :
   $(a_n^T w - b_1^T w)^2 + (a_n^T w - b_2^T w)^2 + ... + (a_n^T w - b_n^T w)^2 + \lambda \parallel w \parallel_2^2$

$\frac{df(w)}{w} =$

$2(a_1^T w - b_1^T w)(a_1 - b_1) + 2(a_1^T w - b_2^T w)(a_1 - b_2) + ... + 2(a_1^T w - b_n^T w)(a_1 - b_n)$
$2(a_2^T w - b_1^T w)(a_2 - b_1) + 2(a_2^T w - b_2^T w)(a_2 - b_2) + ... + 2(a_2^T w - b_n^T w)(a_2 - b_n)$
:
:
$2(a_n^T w - b_1^T w)(a_n - b_1) + 2(a_n^T w - b_2^T w)(a_n - b_2) + ... + 2(a_n^T w - b_n^T w)(a_n - b_n) + \frac{d(\lambda \|w\|_2^2)}{dw}$

Solving Lambda Term $\frac{d(\lambda \|w\|_2^2)}{dw}$:

$lambda \parallel w \parallel_2^2 = \lambda((\sum_{k=1}^n w^2)^{1/2})^2)$

$lambda \parallel w \parallel_2^2 = \lambda(\sum_{k=1}^n w^2)$

$\frac{d(\lambda \|w\|_2^2)}{dw} = 2\lambda \sum_{k=1}^n w$

$\frac{d(\lambda \|w\|_2^2)}{dw} = 2\lambda w$

Thus Solving

$\frac{df(w)}{w} = \sum_{i=1}^n \sum_{j=1}^n (a_i^T w - b_i^T w)(a_i - b_j) + 2\lambda w$

Linear Algebra Warm Up

1. (a) $y^T z = (1 \cdot 2) + (3 \cdot 3) = 11$

   (b) $Xy = \begin{bmatrix} (2 \cdot 1) + (4 \cdot 3) \\ (1 \cdot 1) + (3 \cdot 3) \end{bmatrix} = \begin{bmatrix} 14 \\ 10 \end{bmatrix}$

   (c) Yes X is invertible because it is a square matrix and has a determinant which is not equal to 0
   $X^{-1} = \frac{1}{(2 \cdot 3) + (-1 \cdot -4)} \begin{bmatrix} 3 & -4 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1.5 & -2 \\ -0.5 & 1 \end{bmatrix}$

   (d) The rank of a matrix is the maximum number of its linearly independent column vectors or row vectors. This can be found by counting the number of non-zero rows or non-zero columns. Therefore, if we have to find the rank of a matrix, we will transform the given matrix to its row echelon form and then count the number of non-zero rows.
   First we reduce matrix to row echelon form = $\begin{bmatrix} 2 & 4 \\ 0 & 3 \end{bmatrix}$
   Then calculate the number of linearly independent rows = 2
   Thus, Rank of Matrix X = 2

2. **Orthognal**: A square matrix with real numbers or elements is said to be an orthogonal matrix, if its transpose is equal to its inverse matrix. Or we can say, when the product of a square matrix and its transpose gives an identity matrix, then the square matrix is known as an orthogonal matrix

3. **Positive Semi Definite**: Matrix A is a positive semi-definite if it is a Hermitian matrix whose eigen values are non negative

4. **Psuedo Inverse**: The pseudo-inverse is defined and unique for all matrices whose entries are real or complex numbers. It is a generalization of the matrix inverse when a matrix isn't invertible and it does this through singular value decomposition.

Probability and Statistics Recap

1. Sample Mean $= \frac{1+1+0+1+0}{5} = \frac{3}{5}$
2. Sample variance $= S^2$

   $S^2 = \frac{1}{n}\sum_{i=-1}^{n}(x_i - \overline{x})^2$

   $= \frac{1}{5}\left[(\frac{2}{5})^2 + (\frac{2}{5})^2 + (-\frac{3}{5})^2 + (\frac{2}{5})^2 + (-\frac{3}{5})^2\right]$

   $= \frac{1}{5}\left[\frac{3\cdot4}{25} + \frac{2\cdot9}{25}\right]$

   $= \frac{6}{25}$
3. $P(S) = 0.5^5 = \frac{1}{32}$

Machine Learning

1. (a) $Loss(x, y, w) = (\sigma(w.\phi(x)) - y)^2$

   $Loss(x, y, w) = (\frac{1}{1+e^{-w.\phi(x)}} - y)^2$

   (b) $Loss(x, y, w) = (\sigma(w.\phi(x)) - y)^2 = (\frac{1}{1+e^{-w.\phi(x)}} - y)^2$

   Let p $= \sigma(w.\phi(x)) \Rightarrow Loss(x, y, z) = (p - y)^2$

   $\frac{d(Loss(x,y,z))}{dw} = \frac{(p-y)^2}{dw} = 2(p - y)\frac{d(p)}{dw}$

   Solving for $\frac{d(p)}{dw}$:

$$\frac{d(p)}{dw} = \frac{d(\left[1+e^{-w.\phi(x)}\right]^{-1})}{dw} = -1(\frac{1}{1+e^{-w.\phi(x)}})^{-2} \cdot (-\phi(x) \cdot e^{-w.\phi(x)}) \quad ::: \quad \frac{d(e^{ax})}{dx} = ae^{ax}$$

$$\frac{d(p)}{dw} = \phi(x)(\frac{1}{1+e^{-w.\phi(x)}})^{-2} \cdot (e^{-w.\phi(x)}) = \phi(x)\left[\frac{1}{1+e^{-w.\phi(x)}} \cdot \frac{e^{-w.\phi(x)}}{1+e^{-w.\phi(x)}}\right]$$

$$\frac{d(p)}{dw} = \phi(x)\left[\frac{1}{1+e^{-w.\phi(x)}} \cdot \frac{1+e^{(-w.\phi(x))}-1}{1+e^{-w.\phi(x)}}\right] = \phi(x)\left[\frac{1}{1+e^{-w.\phi(x)}} \cdot (1 - \frac{1}{1+e^{-w.\phi(x)}})\right]$$

$$\frac{d(p)}{dw} = \phi(x)p(1-p)$$

Substituting $\frac{d(p)}{dw}$

$$\frac{d(Loss(x,y,z))}{dw} = 2\phi(x)(p-y)p(1-p)$$

(c) We know that the gradient of the loss with respect to w is:

$$\frac{d(Loss(x,y,z))}{dw} = 2\phi(x)(p-y)p(1-p) \text{ , where p} = \frac{1}{1+e^{-w.\phi(x)}}$$

Lets call z = $w.\phi(x)$. If we assume that y = 1 then the value of p that makes the magnitude of the gradient of loss arbitrarily small is p approaching 1. As p approaches 1 the gradient of loss becomes arbitrarily small because of (p-y) and (1-p). We know that p is the sigmoid function which depends on the value of z. As z increases to an arbitrarily large number p approaches 1. And since z is directly proportional to w, that means as the value of w increases to an arbitrarily large number so does z, causing p to approach 1 which causes the gradient of the loss function to become arbitrarily small.Thus the condition for w would be to become arbitrarily large to make the magnitude extremely small.This value of the gradient of the loss function can go extremely small as the magnitude of w increases but can not become 0 as the sigmoid function has two horizontle asymptotes at y=0 and y=1. So p can approach 1 but can never equal 1 thus the gradient can never be exactly 0.

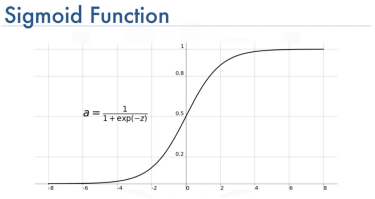

Figure 1: Image of the Sigmoid Function.

2. True

3. (a) Points:
$\phi(x_1) = \begin{bmatrix}1, 0\end{bmatrix}$
$\phi(x_2) = \begin{bmatrix}1, 2\end{bmatrix}$
$\phi(x_3) = \begin{bmatrix}3, 0\end{bmatrix}$
$\phi(x_4) = \begin{bmatrix}2, 2\end{bmatrix}$

Running with First Centres:

First Run ...
Assign Members With centres $\mu_1 = \begin{bmatrix}2, 3\end{bmatrix}$ and $\mu_2 = \begin{bmatrix}2, -1\end{bmatrix}$

Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 10$ and $\mu_2 = 2$
Assigned to: Group 2
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 2$ and $\mu_2 = 10$
Assigned to: Group 1
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 10$ and $\mu_2 = 2$
Assigned to: Group 2
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 1$ and $\mu_2 = 9$
Assigned to: Group 1

Update Centers
$\mu_1 = average(\phi(x_2, \phi(x_4) = \begin{bmatrix}1.5, 2.0\end{bmatrix}$
$\mu_2 = average(\phi(x_1, \phi(x_3) = \begin{bmatrix}2.0, 0.0\end{bmatrix}$

Second Run ...
Assign Members With centres $\mu_1 = \begin{bmatrix}1.5, 2.0\end{bmatrix}$ and $\mu_2 = \begin{bmatrix}2.0, 0.0\end{bmatrix}$

Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 4.25$ and $\mu_2 = 1$
Assigned to: Group 2
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 0.25$ and $\mu_2 = 5$
Assigned to: Group 1
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 6.25$ and $\mu_2 = 1$
Assigned to: Group 2
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 0.25$ and $\mu_2 = 4$
Assigned to: Group 1

Final Centers
$\mu_1 = average(\phi(x_2, \phi(x_4) = \begin{bmatrix}1.5, 2.0\end{bmatrix}$
$\mu_2 = average(\phi(x_1, \phi(x_3) = \begin{bmatrix}2.0, 0.0\end{bmatrix}$

Running with Second Centres:

First Run ...
Assign Members With centres $\mu_1 = \begin{bmatrix}0, 1\end{bmatrix}$ and $\mu_2 = \begin{bmatrix}3, 2\end{bmatrix}$

uclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 2$ and $\mu_2 = 8$
Assigned to: Group 1
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 2$ and $\mu_2 = 4$
Assigned to: Group 1
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 10$ and $\mu_2 = 4$
Assigned to: Group 2
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 5$ and $\mu_2 = 1$
Assigned to: Group 2

Update Centers
$\mu_1 = average(\phi(x_2, \phi(x_4) = \begin{bmatrix}1.0, 1.0\end{bmatrix}$
$\mu_2 = average(\phi(x_1, \phi(x_3) = \begin{bmatrix}2.5, 1.0\end{bmatrix}$

Second Run ...
Assign Members With centres $\mu_1 = \begin{bmatrix}1.0, 1.0\end{bmatrix}$ and $\mu_2 = \begin{bmatrix}2.5, 1.0\end{bmatrix}$

Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 1$ and $\mu_2 = 3.25$
Assigned to: Group 1
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 1$ and $\mu_2 = 3.25$
Assigned to: Group 1
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 5$ and $\mu_2 = 1.25$
Assigned to: Group 2
Euclidean Distance Squared for $\phi(x_1)$ to $\mu_1 = 2$ and $\mu_2 = 1.25$
Assigned to: Group 2

Final Centers
$\mu_1 = average(\phi(x_2, \phi(x_4) = \begin{bmatrix}1.0, 1.0\end{bmatrix}$
$\mu_2 = average(\phi(x_1, \phi(x_3) = \begin{bmatrix}2.5, 1.0\end{bmatrix}$

(b) Running K-Means multiple times on the same dataset with different random initializations gives a higher probability of converging to a global minima rather than doing it with one initialization which most probably converges to local minima.

(c) No, because K Means is very sensitive to scaling and will give more weightage to the features with the larger variances. This will cause the larger variance feature

to be more affected by the Euclidean distance causing the assignment of each cluster to be affected by the larger variance feature. Thus giving different clusters before and after scaling. This would also be true if we only scaled only dimensions as well.