AI-701: Artificial Intelligence **Assignment # 01**

Course Examiner & Lecturer: Dr. Muhammad Haris
Email: muhammad.haris@mbzuai.ac.ae

Name & ID:.....

Course ID: AI - 701 Term: Fall 2021

Due Date: 21^{st} September, 2021 (23:59)

1 AI and Intelligent Agents (5+5+5+6+3) = 24

1. Define in your own words: (a) intelligence, (b) artificial intelligence, (c) agent, (d) rationality, and (e) logical reasoning.
2. Many of the computational models of cognitive activities that have been proposed involve quite complex mathematical operations, such as convolving an image with a Gaussian or finding a minimum of the entropy function. Most humans (and certainly all animals) never learn this kind of mathematics at all, almost no one learns it before college, and almost no one can compute the convolution of a function with a Gaussian in their head. What sense does it make to say that the "vision system" is doing this kind of mathematics, whereas the actual person has no idea how to do it?
3. Various subfields of AI have held contests by defining a standard task and inviting researchers to do their best. Examples include the DARPA Grand Challenge for robotic cars, The International Planning Competition, the Robocup robotic soccer league, the TREC information retrieval event, and contests in machine translation, speech recognition. Investigate five of these contests, and describe the progress made over the years. To what degree have the contests advanced the state-of-the-art in AI? Do what degree do they hurt the field by drawing energy away from new ideas?
4. Define in your own words the following terms: agent, agent function, agent program, rationality, autonomy, and reflex agent.
5. For each of the following activities, give a PEAS description of the task environment: (a) medical diagnosis system, (b) satellite image analysis system, and (c) interactive English tutor.

2 Calculus Recap (1+3+6)=10

1. If $y = x^3 + x - 5$ then what is the derivative of y with respect to x?
2. Consider the following quadratic function: $f(\theta) = 1/2 \sum_{i=1}^{n} w_i(\theta - x_i)^2$, where $x_1, ..., x_n$ are points in 1D space taking on real values and $w_1, .., w_n$ are positive real numbers denoting the importance of each of these points. What value of θ minimizes $f(\theta)$?
3. Compute the gradient of the following function:$f(w) = \sum_{i=1}^{n} \sum_{j=1}^{n} (a_i^T w - b_j^T w)^2 + \lambda ||w||_2^2$. Where $w \in \mathbb{R}^d$ is represented as column vector, $a_i, b_j \in \mathbb{R}^d$ are in the form of column vectors and $\lambda \in \mathbb{R}, w = (w_1, ...w_d)^T \in \mathbb{R}$ and $|| \cdot ||_2$ is the L_2 norm.

3 Linear Algebra Warmup (4+2+2+2) = 10

1. Vectors and Matrices:
   Consider the matrix X and the vectors y and z below:

$$X = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} y = \begin{bmatrix} 1 \\ 3 \end{bmatrix} z = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

   (a) What is the inner product of the vectors y and z? (this is also sometimes called the dot product, and is sometimes written $y^T z$)

(b) What is the product Xy?

(c) Is X invertible? If so, give the inverse, and if no, explain why not.

(d) What is the rank of X? Explain your answer.

2. What does it mean that matrix A is orthonormal?

3. What does it mean that matrix A is positive semi-definite?

4. A is a matrix. What is the definition of its pseudo-inverse?

4 Probability & Statistics Recap (1+2+3) = 6

1. Consider a sample of data $S = \{1, 1, 0, 1, 0\}$ created by flipping a coin x five times, where 0 denotes that the coin turned up heads and 1 denotes that it turned up tails.

   (a) What is the sample mean for this data?

   (b) What is the sample variance for this data?

   (c) What is the probability of observing this data, assuming it was generated by flipping a coin with an equal probability of heads and tails (i.e. the probability distribution is $p(x = 1) = 0.5, p(x = 0) = 0.5$).

5 Machine Learning (1+4+4+1+3+2+3) = 18

1. We want to predict numeric rating for movies reviews. We will use a non-linear predictor, it takes a movie review x and returns $\sigma(w.\phi(x))$, where $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic regression. This function squashes a real number to the range (0,1). We will employ squared loss here. Assume that movie rating y real-valued variable in the range [0,1].

   (a) Write the expression for $Loss(x, y, w)$ for a single data point (x,y).

   (b) Compute the gradient of the loss with respect to w.

   (c) Suppose there is one data point (x, y) with some arbitrary $\phi(x)$ and y = 1. List conditions for w to make the magnitude of the gradient of the loss with respect to w arbitrarily small (minimize the magnitude of the gradient)? If so, how small? Can the magnitude of the gradient with respect to w ever be exactly zero? You are allowed to make the magnitude of w arbitrarily large but not infinity. Why we're interested in the magnitude of the gradients is because it governs how far gradient descent will step. As an example, if the gradient is close to zero when w is very far from the optimum, then it could take a long time for gradient descent to reach the optimum (if at all). This is known as the *vanishing gradient problem* when training neural networks.

2. True or False: Every Boolean function can be represented by a network with a single hidden layer.

3. Suppose we have a feature extractor ϕ that produces 2-dimensional feature vectors, and a toy dataset $\mathcal{D}\_train = \{x_1, x_2, x_3, x_4\}$ with
   $\phi(x_1) = [1, 0]$
   $\phi(x_2) = [1, 2]$
   $\phi(x_3) = [3, 0]$
   $\phi(x_4) = [2, 2]$

   (a) Run 2-means on this dataset until convergence. Please show your work. What are the final cluster assignments z and cluster centers μ ? Run this algorithm twice with the following initial centers:
   $\mu_1 = [2, 3]$ and $\mu_2 = [2, -1]$
   $\mu_1 = [0, 1]$ and $\mu_2 = [3, 2]$

(b) What is the advantage of running K-means multiple times on the same dataset with the same K, but different random initializations?

(c) If we scale all dimensions in our initial centroids and data points by some factor, are we guaranteed to retrieve the same clusters after running k-means (i.e will the same data points belong to the same cluster before and after scaling)? What if we scale only certain dimensions? If your answer is yes, provide a short explanation. If it is no, provide a counterexample.
