

CS 771A: Introduction to Machine Learning			Quiz 4 (01 Nov 2019)	
Name				30 marks Page 1 of 2
Roll No		Dept.		

Instructions:

1. This question paper contains 1 page (2 sides of paper). Please verify.
2. Write your name, roll number, department above in **block letters neatly with ink**.
3. Write your final answers neatly **with a blue/black pen**. Pencil marks may get smudged.
4. Don't overwrite/scratch answers especially in MCQ. We will entertain no requests for leniency.
5. Do not rush to fill in answers. You have enough time to solve this quiz.

Q1. Write T or F for True/False (write **only in the box on the right hand side) (8x2=16 marks)**

1	The Adagrad method is a technique for choosing an appropriate batch size when training a deep network.	
2	The largest value the Gaussian kernel can take on any two points depends on the value of the bandwidth parameter used within the kernel.	
3	k-means++ initialization is one of the algorithms that cannot be kernelized easily since it involves probabilities and sampling.	
4	Suppose G is the Gram matrix of n data points $\mathbf{x}^1, \dots, \mathbf{x}^n \in \mathbb{R}^2$ with respect to the homogeneous polynomial kernel of degree $p = 2$. Then G must be pos. semi def.	
5	If for some \mathbf{w}^* we have $y^i = \langle \mathbf{w}^*, \mathbf{x}^i \rangle, i \in [n]$ then kernel regression with $K(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + 1)^2$ cannot get zero training error w.r.t least squares loss on this data	
6	Kernel k-means clustering with the quadratic kernel results in a larger model size than what is possible if we had done linear k-means (i.e. with the linear kernel).	
7	A NN with a single hidden layer and a single output node with all nodes except input layer nodes using ReLU activation will always learn a differentiable function.	
8	Dropout is a technique that takes a training set and randomly drops training points to reduce the training set size so that training can be done faster	

Q2. Suppose we have n distinct data points data points $\mathbf{x}^1, \dots, \mathbf{x}^n \in \mathbb{R}^2$. Consider the Gram matrix G w.r.t the Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \cdot \|\mathbf{x} - \mathbf{y}\|_2^2)$. Answer in the boxes only. (6 marks)

2.1	Write down the value of $\text{trace}(G)$ as $\gamma \rightarrow 0$	
2.2	Write down the value of $\text{trace}(G)$ as $\gamma \rightarrow \infty$	
2.3	Write down the value of $\text{rank}(G)$ as $\gamma \rightarrow 0$	
2.4	Write down the value of $\text{rank}(G)$ as $\gamma \rightarrow \infty$	
2.5	If instead of being distinct, had all the points been the same i.e. $\mathbf{x}^1 = \mathbf{x}^2 = \dots = \mathbf{x}^n$, write down the value of $\text{rank}(G)$ as $\gamma \rightarrow 0$	
2.6	If instead of being distinct, had all the points been the same i.e. $\mathbf{x}^1 = \mathbf{x}^2 = \dots = \mathbf{x}^n$, write down the value of $\text{rank}(G)$ as $\gamma \rightarrow \infty$	

Q3 Let $\mathbf{x} = [1, -1]^T, \mathbf{y} = [-1, 1]^T \in \mathbb{R}^2$. Define the function $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as $f(\mathbf{z}) = z_1 \cdot \mathbf{x} + z_2 \cdot \mathbf{y}$ for any $\mathbf{z} = [z_1, z_2] \in \mathbb{R}^2$. Define another function $g: \mathbb{R} \rightarrow \mathbb{R}^2$ as $g(r) = [r, r^2]$ where $r \in \mathbb{R}$. Let $h: \mathbb{R} \rightarrow \mathbb{R}^2$ be defined as $h(r) = f(g(r))$. Derive a general expression for $\frac{dh}{dr}$ using the chain rule giving major steps of derivation and then evaluate $\frac{dh}{dr}$ at $r = 3$. **(6 + 2 = 8 marks)**

----- END OF QUIZ -----

ROUGH WORK
Nothing written here will get graded