UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**COMP0081**
ASSESSMENT PATTERN	:	**COMP0081A7PC**
MODULE NAME	:	**COMP0081 - Applied Machine Learning**
LEVEL:	:	**Postgraduate**
DATE	:	**10/05/2021**
TIME	:	**14:30**

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

Year
2020/21

Additional material	N/A
Special instructions	N/A
Exam paper word count	1000

TURN OVER

COMP0081: Applied Machine Learning

Final Exam (Main examination period)

For 2020–2021 Cohort, all levels

Always provide justification and show any intermediate work for your answers. A correct but unsupported answer may not receive any marks.

The exam includes 6 questions in multiple parts, worth a total of 120 marks. All questions must be answered. The marks available for each question are indicated in the square brackets.

1. **True/False**

Please identify if statements are either True or False. Please justify your answer **if false**. Correct "True" questions yield 2 points. Correct "False" questions yield one point for the answer and one point for the justification.

(a) **(T/F)** A learned kernel SVM model (with RBF kernel) requires you to store some of the training data. *[2 marks]*

(b) **(T/F)** Random forests is one of the few machine learning algorithms that makes no assumptions on the data. *[2 marks]*

(c) **(T/F)** K-means algorithm can never output two clusters such that the distance between the centers of the two clusters is shorter than the diameters of any one of the clusters. *[2 marks]*

(d) **(T/F)** The curse of dimensionality tells us that if we sample n data points uniformly at random within a hypercube of dimensionality d, all pairwise distances converge as $n \to \infty$. *[2 marks]*

(e) **(T/F)** The optimization of deep neural networks is a convex minimization problem. *[2 marks]*

(f) **(T/F)** Any real symmetric matrix is a well-defined kernel. *[2 marks]*

(g) **(T/F)** As the validation set becomes extremely large, the validation error approaches the test error. *[2 marks]*

(h) **(T/F)** PCA is always better than random projection for dimensionality reduction. *[2 marks]*

(i) **(T/F)** In general, models that are more easily interpretable often are more accurate. *[2 marks]*

(j) **(T/F)** The L1 regularizer encourages sparse solutions. *[2 marks]*

(k) **(T/F)** Random Forests are bagged decision trees with one additional modification: Each splitting dimension is chosen completely uniformly at random. *[2 marks]*

(l) **(T/F)** After convergence, the K-means cluster centers are always original data points. *[2 marks]*

(m) **(T/F)** A classical way to make a quantity differentially private is to add random noise drawn from a symmetric distribution to that quantity. *[2 marks]*

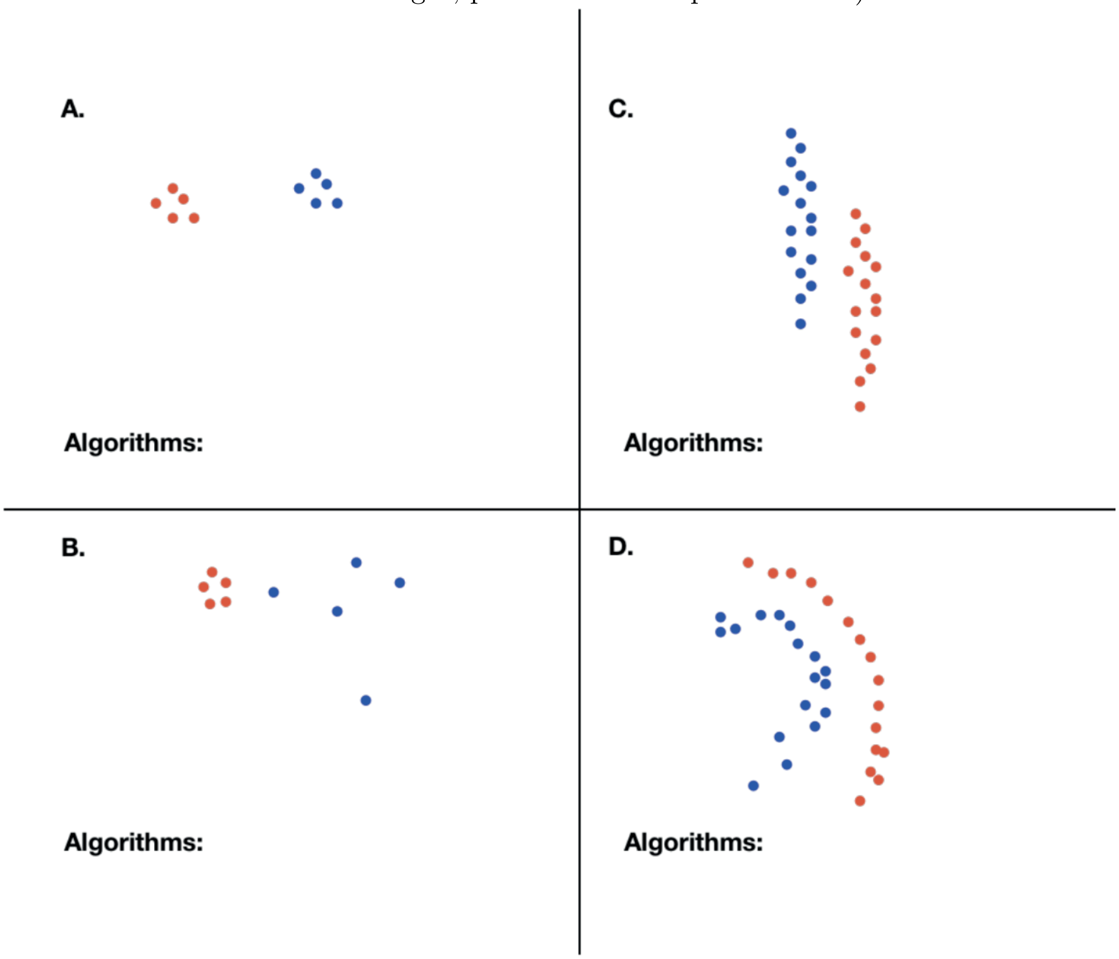
(n) **(T/F)** If a classifier obtains 0% training error it cannot have 100% testing error. *[2 marks]*

(o) **(T/F)** As your training data set size, n, approaches infinity, the k-nearest neighbor classifier is guaranteed to have an error no worse than twice the Bayes optimal error. *[2 marks]*

[30 marks total]

2. **Which Clustering Algorithm?**

For each of the following clustering of (2 dimensional) points into two clusters (i.e all run with $K = 2$) (i.e., red points for cluster 1 and blue points for cluster 2), write down which amongst: **k-means, single linkage clustering, Gaussian mixture models, or none of the three** could have led to the cluster assignments shown. (more than one answer could be right, put down all the possible ones).



[20 marks total]

3. **Classifiers**

(a) Your Decision Tree classifier has a training error of 0% and a testing error of 92%. What are two possible interventions which can reduce the testing error? *[2 marks]*

(b) For k-fold cross validation, describe the positive and negative effects as $k \to n$. When would you be most inclined to use $k = n$? *[3 marks]*

(c) For the following pairs of algorithms and hyperparameters, would *increasing* the parameter help reduce overfitting? Yes or no.

   i. Neural networks: number of hidden units

   ii. Decision trees: maximum depth

   iii. Logistic Regression, with $\lambda\|\mathbf{w}\|_2^2$ penalty: the value of λ

   iv. Boosting: the number of iterations T

[4 marks]

(d) Name two advantages of decision trees over nearest neighbors *[2 marks]*

(e) Imagine 10% of your binary training data are accidentally mis-labelled. What training error will AdaBoost converge to after enough rounds of boosting? (Assume all your inputs are unique.) *[2 marks]*

(f) Describe a data scenario in which AdaBoost is not a good choice. Justify your answer. *[4 marks]*

(g) Given a distribution P you can sample a training set $\mathcal{D}$ and obtain a classifier h. Imagine you train m such classifiers $h_1, ..., h_m$ on m data sets $\mathcal{D}_1, ..., \mathcal{D}_m$, each drawn i.i.d. from the data distribution P . As you increase m from $m = 1$ to $m \gg 0$,

   i. Show how you can use these classifiers to define a new, lower-variance classifier $\hat{h}$

   ii. What happens to the variance of $\hat{h}$ in the limit as $m \to \infty$?

   iii. How does the bias of $\hat{h}$ compare to the bias of h?

[3 marks]

[20 marks total]

4. **Optimization and Regularization**

(a) Explain the trade-offs in using stochastic (mini-batch) and pure gradient descent. When would you select one over the other and why? *[2 marks]*

(b) Suppose you want to avoid overfitting and implement **early stopping** for your model. You train the model using stochastic gradient descent, with mini-batch updates. After every batch you check the validation loss, and stop as soon as it goes up. Is this a good idea? Why? *[3 marks]*

(c) Suppose you have a linear regression problem N training examples $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^D$. Consider a linear model with a squared error loss and L2 regularization, known as *ridge regression*. This model has a loss function

$$\mathcal{L}(\mathbf{x}_{1:N}, y_{1:N}) = \sum_{i=1}^{N}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}.$$

i. For a fixed value of the regularization parameter λ, this loss function has an analytic solution for $\mathbf{w}$. What is it? (Show your derivation!) *[5 marks]*

ii. What might be the reasons for using gradient-based optimization rather than the equation you just derived? *[2 marks]*

iii. Can you optimize the parameter λ by gradient descent? Why or why not? Explain how you would best estimate λ. *[3 marks]*

[15 marks total]

5. **Privacy of the Median and the Mean**

In class we looked at the idea of differential privacy. The idea in differential privacy was that we considered a randomized algorithm A to be an ϵ-differentially private algorithm if for every set C and every sample S, S' that differ on at most one point,

$$P(A(S) \in C) \leq e^{\epsilon} P(A(S') \in C)$$

This notion of privacy is rather too stringent as we require the above to hold for **arbitrary** S and S' that differ by at most one. However, naturally occurring samples S might be easier to privatize.

In this problem we will explore the **median** and **mean** and how much private information these quantities each reveal.

Questions:

(a) If we want to use the Laplace mechanism, that is add Laplace distributed noise to our quantity $f(\mathcal{X})$ before releasing it. **What is the amount of Laplace noise we need to add to release the mean of n points $\mathcal{X} = \{x_1, \ldots, x_n\}$ where each x_i is either 0 or 1?** Recall that Laplace mechanism adds to the quantity $f(\mathcal{X})$ Laplace distributed noise whose density function is given by $p(x; b) = \frac{1}{2b} \exp(-|x|/b)$. So **the question is asking you to specify parameter b as a function of** ϵ so that if $f(\mathcal{X})$ is the **mean** then the final noisy quantity $f(\mathcal{X}) + \eta$, where $\eta \sim p(x; b)$, is ϵ-differentially private via the Laplace mechanism. *[5 marks]*

(b) Next, if we want to release the **median** of the same binary $x_1, \ldots, x_n$ is a differentially private way, **how much Laplace noise do we need to add?** (what is the parameter b) *[5 marks]*

(c) Now consider the case when we are interested not in the stringent notion of differential privacy but rather we want to ask the question, for the *given dataset* S, how much noise do we need to add (i.e., this noise is specific to sample S) to make the algorithms private. Assume that $n > 10$ for this problem. Specifically, consider the case when $S = \{x_1, \ldots, x_n\}$ are such that $n/3$ of the x's are ones and the rest are 0's. Now for this particular sample S, if we want that for any S' that is different from S by, at most one point, how much noise do we need to add to the **median** to ensure that for a given ϵ, $P(A(S) \in C) \leq e^{\epsilon} P(A(S') \in C)$? For this sample, how much noise do we need to add to the **mean** to ensure that for a given ϵ, $P(A(S) \in C) \leq e^{\epsilon} P(A(S') \in C)$? *[10 marks]*

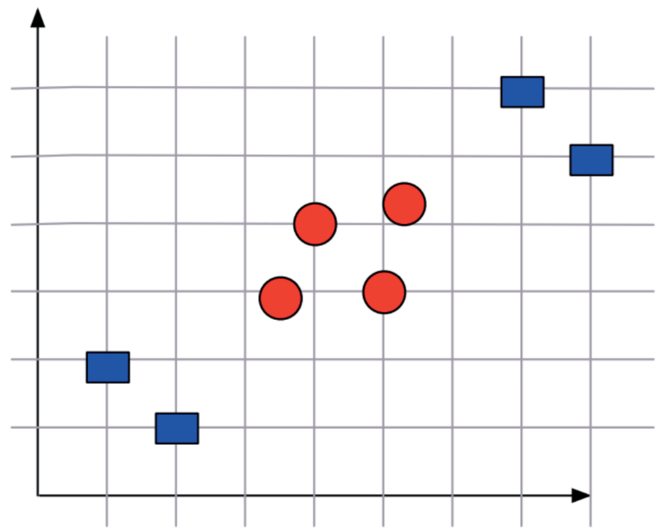
[20 marks total]

6. **Kernel Methods**

(a) Name one condition that is necessary and sufficient for a matrix $\mathbf{K}$ to be positive semi-definite. *[2 marks]*

(b) Which of the following algorithms can be kernelized: a) Decision Trees, b)Linear Regression, c) Gaussian Processes. Justify your answer. *[3 marks]*

(c) Consider the following data set. Draw a plausible decision boundary for a hard-margin SVM with polynomial kernel.



[2 marks]

(d) Let m be the number of support vectors of an SVM trained on n data points (with RBF kernel). For a fixed n imagine you increase the dimensionality d of the data until it becomes very large. How would you expect the ratio $\frac{m}{n}$ to change as $d \gg 0$? *[3 marks]*

(e) Describe a scenario in which you may want to use a kernel SVM with linear kernel instead of a standard linear (primal) SVM. *[3 marks]*

(f) You are given a non-linear regression data set. You are deciding between training a Gaussian Process or kernelized linear regression (both with RBF Kernel). Which one will have lower testing / training error? *[2 marks]*

[15 marks total]

Question	Points Scored	Max Points
True/False		30
Which Clustering Algorithm?		20
Classifiers		20
Optimization and Regularization		15
Privacy of the Median and the Mean		20
Kernel Methods		15
TOTAL		120

END OF PAPER