COMPSCI 4ML3-C01
Instructor: Hassan Ashtiani
Midterm Exam
March 2023

First Name: ______________________

Last Name: ______________________

Student Number ______________________

Note that

- The time limit is 90 Minutes.
- This examination paper includes 8 pages (including this cover page and the last blank page) and 9 questions. You are responsible for ensuring that your copy of the papers is complete. Bring any discrepancy to the attention of your invigilator.
- Total of points is 100
- If you are not sure about the meaning of a question, use your own judgment to make the best assumption.

Special Instructions:

1. The exam is closed-book. No crib sheets are allowed. You are allowed to use Standard McMaster calculator.

Grade Table

Question	Points	Score
1	5	
2	5	
3	5	
4	5	
5	5	
6	5	
7	5	
8	20	
9	45	
Total:	100	

1. (5 points) Consider the 1-dimensional ordinary least squares problem, where we fit $ax+b$ to our data. Assume our data set is centered, in the sense that the average value of x over the data set is zero. What will be optimal value for b? (choose only one option)

   - ◯ 0
   - ◯ 1
   - ◯ average value of y over the data set
   - ◯ covariance of x and y (over the data set)

2. (5 points) We are using ordinary least squares and observe that $X^T X$ is not invertible. Which conclusion is valid? Choose only one option (assume N is the number of data points and d is the number of dimensions)

   - ◯ The least squares problem has a unique solution
   - ◯ The least squares problem does not have any solutions
   - ◯ $Rank(X) < d$
   - ◯ $Rank(X) < N$

3. (5 points) Which choice is often **incorrect** about changing the dimensionality of the data? Choose only one option (here, the downstream task just means the task that we want to do after changing the dimension)

   - ◯ Mapping the data into a lower dimensional space (e.g, by using PCA) can help with the overfitting issue for a downstream supervised learning task (e.g., regression).
   - ◯ Mapping the data into a higher dimensional space (e.g, by using polynomial mappings) can help with the overfitting issue for a downstream supervised learning task (e.g., regression).
   - ◯ Mapping the data into a higher dimensional space (e.g, by using polynomial mappings) can can increase the computational cost of a downstream supervised learning task (e.g., regression).
   - ◯ Mapping the data into a lower dimensional space (e.g., by using PCA) can increase the risk of losing some information.

4. (5 points) Which one is NOT a common approach for dealing with overfitting? Choose only one.
   - ◯ Using the kernel trick
   - ◯ Adding more training data
   - ◯ Using regularization
   - ◯ Reducing the dimensionality of the data

5. (5 points) Assume we have already found the solution for kernel least squares for a data set of size N, and now would like to estimate the value of y for a new x (x is a test point). Assume computing kernel of x takes $O(d)$ time; in other words computing $k(x, z)$ for two points takes $O(d)$ time. What is the time complexity of computing the estimated $\hat{y}$ at point x?
   - ◯ $O(N)$
   - ◯ $O(d)$
   - ◯ $O(Nd)$
   - ◯ $O(N + d)$

6. (5 points) What is the difference between using polynomial mappings (basis functions) for least-squares versus using the polynomial kernel?
   - ◯ kernel least squares is often faster when degree of the polynomial is large
   - ◯ kernel least squares is often more accurate when degree of the polynomial is large
   - ◯ kernel least squares is often faster when the number of data points is very large.
   - ◯ kernel least squares is often more accurate when the number of data points is very large.

7. (5 points) Which of these is computationally more expensive? Choose only one.
   - ◯ Finding a line with minimum classification error on a data set when the number of data points is large but the number of dimensions is moderate.
   - ◯ Finding a line with minimum classification error on a data set when the number of dimensions is high but the number of data points is moderate.
   - ◯ Finding the maximum margin classifier when the number of data points is large but the number of dimensions is moderate.
   - ◯ Finding the maximum margin classifier when the number of dimensions is high but the number of data points is moderate.

8. (20 points) Assume that the latency of the google website follows the exponential distribution. Recall that the exponential distribution is defined over $[0, \infty)$ and has the probability density function $f(t) = \lambda e^{-\lambda t}$. We are interested to estimate the value of λ based the maximum likelihood principle. For this, we have collected N independent samples $t_1, t_2, \ldots t_N$ from the exponential distribution (i.e., N latencies). Find the maximum likelihood estimate of λ based on $t_1, t_2, \ldots t_N$. Show your work and simplify your answer as much as possible.

9. (a) (20 points) Consider the real-valued function $f : \mathbb{R} \to \mathbb{R}$ where

$$f(x) = \begin{cases} 0 & x \leq 1 \\ 1 & x > 1 \end{cases}$$

We want to fit a line with zero y-intercept ($\hat{y} = ax$) to the function $y = f(x)$. Our goal is to find the parameter a such that expected squared error is minimized (i.e., least squares):

$$a^{LS} = \arg\min_{a \in \mathbb{R}} \mathbb{E}_{x \sim P_x} \left(f(x) - ax\right)^2$$

Here, P_x is the distribution of x, and is assumed to be a uniform distribution over $[0, 2]$. Compute the value of a^{LS}. Show your work and simplify the final answer.

(b) (5 points) What is the expected squared error of a^{LS}? Show your work and simplify your final answer.

(c) (20 points) Since $f(x)$ only outputs values in $\{0, 1\}$, one can look at the problem as a binary classification task too. With a slight change, consider the linear classifier parameterized by a and b:

$$\hat{y} = sign(ax + b) = \begin{cases} 1 & ax + b \geq 0 \\ 0 & ax + b < 0 \end{cases}$$

Here, the goal is to minimize the classification error instead of the squared loss:

$$\arg\min_{a,b \in \mathbb{R}} \mathbb{E}_{x \sim P_x} \ell^{0-1}\left(f(x), sign(ax + b)\right)$$

where ℓ^{0-1} is simply the 0-1 classification loss:

$$\ell^{0-1}\left(y, \hat{y}\right) = \begin{cases} 0 & y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

Consider the same distribution P_x as before. Which values of a and b minimize the expected classification error? What will be the expected error in this case (i.e., with the best choice of a and b)?

You can use this page as scrap paper. Return this page with the rest of the papers.

The End