COMP5328 Sample Final Exam

This is a sample final exam of COMP5328.

The information about the final exam of COMP5328 is provided below.

Duration: 3 hours and 10 minutes (190 minutes). This includes 10 minutes of reading time, but you can start writing whenever you are ready.

- **Buffer time (Assignment)**: You will be allowed a buffer time of 15 minutes. This means that you have 15 minutes after the 'Due' date and time to still be able to submit your exam before the assignment closes. Please note that your assignment will be marked as a late submission if you submit after the 'Due' date and time. If you are unable to submit your exam within 15 minutes of buffer time, you should apply for special consideration. Buffer time does NOT mean you have extra time to complete your exam.
- **Upload time**: The final exam has 15 minutes of upload time added to the duration to allow you to upload images/workings etc. as per your exam instructions. Do NOT treat this as extra time. The upload time must be used solely to save and upload your documents correctly as per the exam instructions.

Format: The final exam is a take-home exam with 10 questions. You should read the questions, and FILL WORDS /INSERT IMAGES OF YOUR HAND-WRITTEN ANSWERS IN THE TEX TEMPLATE. For hand-written answers, write them down on A4 paper clearly. Take images just covering the entire A4 paper. Only the image formats of ".png", and ".jpeg" are accepted. You should attempt all questions and follow the instructions for each question carefully.

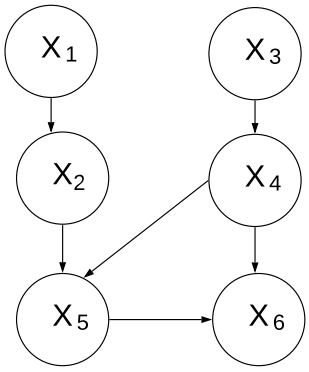
	Question type	Points	Recommended time spent
Part A	You are allowed to answer the questions in text only	40	60
Part B	You are allowed to answer the questions in text and/or image. Words on the images are only to explain the formula, the derivation, or the logic chain. If you want to explain concepts/statements by words, please type the words in the latex answer template.	60	120

Part A: You are allowed to answer the questions in text only.

Question 1 (Approximation error, estimation error) [10pts]
Suppose we have a task which is to identify different persons and a training dataset which contains pairs of human face image and identity. Two predefined hypothesis classes are given, which are linear classifiers and deep neural networks, respectively.
1). Which hypothesis class has a larger Vapnik–Chervonenkis (VC) dimension? Explain why in details.
2). Which hypothesis class is more possible to produce a smaller approximation error? Explain why in details.
3). Which hypothesis class is more possible to produce a smaller estimation error when trained with this dataset? Explain why in details.
Answer:
1) Deep neural networks class will have a larger VC dimension. By definition, VC dimension explain the largest data size which can be shattered by the hypothesis class. It can measure the complexity of the hypothesis class. So the deep neural network is much more complex than the linear class.

2) Deep neural network. Since it contains more types of hypothesis so there is a higher chance that the best hypothesis in the universal function space ($c = \text{argmin}_h R_h$) equals the optimal hypothesis in the predefined hypothesis class ($h^* = \text{argmin}_h R_h$).
3) Linear classifier. Given the sample size fixed, linear classifier is easier to learn compared to the deep neural network.

Question 2 (Causal Inference) [10pts]
We have a directed graphic causal model as follows.



[Note that you can use X-1, X-2, X-3,…,X-4, X-5 and X-6 to refer to X_1, X_2 X_3, X_4, X_5, and X_6, respectively. You can use NULL to denote an empty set $\{\emptyset\}$.]
1). Let $\boldsymbol{X} = \{X_2\}$ and $\boldsymbol{Y} = \{X_6\}$, Provide a set which d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$. Explain why in details.
2). Let $\boldsymbol{X} = \{X_1, X_2\}$ and $\boldsymbol{Y} = \{X_3, X_4\}$, Provide a set d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$. Explain why in details.

Answer:
1) S = $\{X_5\}$. For the chain $X_2 \to X_5 \to X_6$, since the middle node X_5 in S, so S d-separates X and Y.
2) S = $\{\}$

Part B: You are allowed to answer the questions in text and/or image.

Question 5 (Sparse coding) [10pts]
In sparse coding, we prefer sparse matrices.
1). Please briefly explain the concept of sparsity.
2). The definition of the l_p norm is as follows:

$$\ell_p = \|\alpha\|_p = (\sum_{j=1}^{k} |\alpha_j|^p)^{1/p},$$

where $\alpha \in \mathbb{R}^k$. Among ℓ_0, ℓ_1, ℓ_2 norms, which one is the best measure of sparsity? Please give a brief explanation.

3). Explain why the ℓ_0 norm is hard to be optimized. How to handle the problem?

Question 6 (Expected Risk and Empirical Risk) [10pts]
Let ℓ be a loss function. We can define the expected risk of a hypothesis h as

$$R(h) = \mathbb{E}[\ell(X, Y, h)].$$

And the corresponding empirical risk can be defined as

$$R_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(X, Y, h),$$

where $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is a training sample consists of n pairs of i.i.d. random variables. Let $\mathcal{H}$ be the predefined hypothesis class. We further define

$$f^* = \operatorname*{argmin}_{h \in \mathcal{H}} R(h),$$
$$f_S = \operatorname*{argmin}_{h \in \mathcal{H}} R_S(h).$$

Prove that $R(h_S) - R(h^*) \leq 2 \sup_{h \in \mathcal{H}} [R_S(h) - R(h)]$. Show the detailed proof steps.

Question 7 (Optimization) [10pts]
Suppose we have a function $f(h)$ where h is the hypothesis. By Taylor's theorem, we have

$$f\left(h_{k+1}\right)=f\left(h_k\right)+\eta \nabla f\left(h_k\right)^{\top} d_k+o(\eta).$$

1). If we want to minimize the function value, the steepest gradient descent method tells us that we need to update our hypothesis in the direction of the negative gradient $-\nabla f\left(h_k\right)$. Please explain why.

2). Please explain how Armijo Backtracking line-search works. (Include some mathematical notation and formulas if necessary. You have to carefully explain the meaning of each notation used.)