CITY UNIVERSITY OF HONG KONG

Course code & title	:	CS5489 Machine Learning: Algorithms & Applications
Session	:	Midterm, Semester B 2021
Time allowed	:	Two hours (Feb 23, 7:00pm-9:00pm)

This is the answer sheet for the CS5489 Midterm.
Put all your answers in this document.

---

*This is an **open-book** midterm.*

Instructions:

- Answer all questions in the accompanying Word document **"CS5489-midterm-2021-answersheet.docx"**.
- The following resources are **allowed** during the final:
  - Videos of CS5489 lectures on Zoom,
  - any material on the CS5489 Canvas page, including lecture notes, tutorials, etc.
  - course textbooks
- Any other resources are **not allowed**, for example
  - internet searches
  - classmates
  - other textbooks
- You should stay on Zoom during the entire exam time in case there are any announcements.
  - If you have any questions, please use the private chat function of Zoom to message Antoni.
- By 9:00pm Feb 23, submit the completed quiz to the "Midterm" Assignment on Canvas.
  - If you have trouble accessing Canvas, then you can send the completed docx via email to Antoni (abchan@cityu.edu.hk).

Below is the statement of academic honesty. Read it and put your Name, EID, and student ID to acknowledge that you agree with it and will follow its terms.

Statement of Academic Honesty

I pledge that the answers in this exam are my own and that I will not seek or obtain an unfair advantage in producing these answers. Specifically,

- ❖ *I will not plagiarize (copy without citation) from any source;*
- ❖ *I will not communicate or attempt to communicate with any other person during the exam; neither will I give or attempt to give assistance to another student taking the exam; and*
- ❖ *I will use only approved devices (e.g., calculators) and/or approved device models.*
- ❖ *I understand that any act of academic dishonesty can lead to disciplinary action.*

I pledge to follow the Rules on Academic Honesty and understand that violations may led to severe penalties.

Name: <PUT YOUR NAME HERE>
EID: <PUT YOUR EID HERE>
Student ID: <PUT YOUR STUDENT ID HERE>

Multiple Choice/Selection Questions (30 points)
5 marks each. *For a multiple selection question, an incorrect answer will be penalized 5/K marks, where K is the number of correct answers. If more incorrect answers are given than correct answers, the marks will be 0.*

Q1 ANSWER: <ANSWER> B, E

Q2 ANSWER: <ANSWER > A, C, D

Q3 ANSWER: <ANSWER > A, D, E

Q4. ANSWER: <ANSWER > D, E

Q5. ANSWER: <ANSWER > A, B, E

Q6. ANSWER: <ANSWER > D

Discussion Questions (70 marks)
10 marks each question.

Q7 ANSWER:
<ANSWER HERE>

- [4 marks] A generative model learns how image features are generated for each class, while a discriminative classifier directly discriminates the classes apart using image features.
    - Discrimative model is easier to setup since we don't need to model the underlying probability distributions.
    - Generative model will be better if there are not a lot of data.
- [6 marks] Intepretation:
    - The discriminative model weights can be interpreted as the "key" differences separating the classes.
    - The generative model CCD can be interpreted as the prototypical features representing the class.

Q8 ANSWER:
<ANSWER HERE>
[3 marks] False.
[7 marks] If the CCDs are overlapped, and there will be a region near the decision surface where samples both classes will appear. In particular, the decision surface is where there is 50% probability a point is in each class. Thus, there is still unavoidable training error.

Q9 ANSWER:
<ANSWER HERE>
Any of the two, or other reasonable reason. [5 marks for each fix]

- 1) Since the training data is imbalance, collect more data (or use data augmentation methods) to enlarge regular email dataset. This can help the classifier to learn more about the normal email class.
- 2) use weights on the classes during training. Since most of emails are predicted as spam, we can assign more weights on regular email so that the classifier focuses more on predicting the regular class correctly.
- 3) change the threshold of the classifier. Usually it is T=0, but we can set it to T<0 to make it predict more regular examples (assuming class +1 is normal email, and -1 is spam).

Q10 ANSWER:
<ANSWER HERE>

- GPR [5 marks]
  - Advantages of GPR: When a region does not have enough data, GPR can output a large uncertainty. So there is diagnostic information if the prediction might be poor.
  - Disadvantages of GPR: 1) The outliers may break GPR. 2) The computation cost is very high since there are 10597 data points if we use a non-linear kernel.
- Random forest [5 marks]
  - Advantages of random forest: The prediction is fast compared to GPR.
  - Disadvantages of random forest: The model is very sensitive to outliers.

Q11 ANSWER:
<ANSWER HERE>
Any 2 advices [5 marks each advice]

- 1) try introducing non-linear variable such as polynomial degree variable. Adding non-linear variable can make the model more complex to be able to fit the data better.
- 2) add feature dimensions. The basic reason might be that there are no informative features, so extracting new types of features will help.
- 3) try non-linear regressors, such as, GPR, Random forest, etc. Perhaps the linear trend of the data is constant, and there are non-linear variations around the mean.

Q12 ANSWER:
<ANSWER HERE>
[3 marks] Adaboost
[7 marks] Adaboost will run faster because each weak learner can be run in parallel. The random forest is based on 10 decision trees, which requires 10 steps traversing the tree. In this case, RF will be 10 times slower than Adaboost.

Q13 ANSWER:
<ANSWER HERE>
Any 3 properties [3.33 marks each]

- There is a margin at $z=1$, similar to SVM, that prefers points to be at least this far from the decision boundary.
- Points beyond the margin ($z>1$) are ignored.
- Points that are misclassified ($z<0$) will have decreasing penalty based on how far they are from the boundary. Thus, this loss should ignore very badly classified outliers. It should be robust to outliers.
- The loss will be difficult to train with since it is not convex.

--- END ---