CITY UNIVERSITY OF HONG KONG

Course code & title	:	CS5489 Machine Learning: Algorithms & Applications
Session	:	Midterm, Semester A 2021
Time allowed	:	Two hours (Oct 28, 1:00pm-3:00pm)

This question paper has 4 pages (including this cover page).

---

1. This paper consists of 13 questions.
2. Answer ALL questions.
3. Write your answers in the accompanying **"CS5489-midterm-2021A-answersheet.docx"**.

---

*This is an **open-book** final exam, see the allowed resources below.*

Instructions:

- Answer all questions in the accompanying Word document **"CS5489-midterm-2021A-answersheet.docx"**.
- The following resources are **allowed** during the final:
  - Videos of CS5489 lectures on Zoom,
  - any **unaltered** material on the CS5489 Canvas page, including lecture notes, tutorials, etc.
  - **unaltered** course textbooks
- Any other resources are **not allowed**, for example
  - internet searches.
  - Classmates.
  - other textbooks, other notes.
  - any text/notes copied into your lecture notes or textbook.
  - translation software.
- 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 3:00pm Oct 28, submit the completed final to the "Final" Assignment on Canvas.
  - If you have trouble accessing Canvas, then you can send the completed docx via email to Antoni (abchan@cityu.edu.hk).

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. Which statements about generative and discriminative classifiers are **not** correct? (select all that apply)
A) New feature dimensions can be added to generative classifiers without re-training the model.
B) Generative classifiers are trained by learning the class-conditional and prior distributions.
C) Discriminative classifiers can learn from unlabelled training data.
D) Generative classifiers highly depend on prior probability distributions and Gaussian distribution is always better than other distributions when using Generative classifiers.
E) Bayes' classifier is a generative classifier, while SVM, logistic regression, AdaBoost, XGBoost and Random Forest are discriminative classifiers.

Q2. Which of the following statements are correct about Naïve Bayes (NB) classifiers? (select all that apply)
A) NB classifiers will minimize the probability of making an error.
B) Because NB classifiers do not model correlations between features, the decision boundaries are always aligned with the axes.
C) NB classifiers do not directly learn the posterior probability of the class.
D) NB classifiers cannot overfit to the training data.
E) The NB classifier accuracy highly depends on the correct selection of the CCD.

Q3. Which statements are true about kernel SVM? (select all that apply)
A) The kernel SVM is equivalent to learning a linear SVM on transformed inputs.
B) The training data is no longer necessary after learning the kernel SVM.
C) Any function that outputs positive values is a valid kernel function.
D) The kernel SVM can be applied to non-vector input data, like strings and sets.
E) The kernel trick can reduce both the memory and computation requirements, compared to an explicit transformation and inner-product.

Q4. For a binary classification problem with input feature dimension $d > 2$, which of the following classifiers **in principle** could learn **both** linear and non-linear classification boundaries on different datasets? (select all that apply)
A) Naïve Bayes with the Gaussian distribution as the CCD.
B) SVM with the RBF kernel.
C) AdaBoost with the decision stump as the weak learner, and the number of weak learners is at least 3.
D) SVM with the polynomial kernel of degree=2, $k(\mathbf{x},\mathbf{x}') = (\mathbf{x}^T\mathbf{x}')^2$.
E) Logistic regression for some specific α.

Q5. What is/are the purpose(s) of using L2 regularization on the weights in linear classifiers? (select all that apply)
A) L2 regularization will encourage some weights to be large.
B) L2 regularization will prevent overfitting.
C) In SVM, L2 regularization is equivalent to minimizing the margin distance.
D) In Logistic Regression, L2 regularization is equivalent to a Gaussian prior distribution on the weights.
E) L2 regularization is only used to stabilize the training, and it has no effects on the weights.

Q6. Which statements about linear regression are correct? (select all that apply)
A) Overfitting is more likely when you have large amount of training data.
B) Some of the weight coefficients will approach zero, but not absolutely zero, when applying very large penalty α in LASSO regression.
C) LASSO can be used for feature selection.
D) Some of the weight coefficients will approach zero, but not absolute zero, when applying very large penalty α in Ridge Regression.
E) OLS is a special case of both LASSO and Ridge regression.

Discussion Questions (70 marks)
10 marks each question.

Q7. Consider a situation where you have trained a linear classification model to diagnose mental disorders patients from symptom checklist data. You have 100,000 training examples, which is sufficient data for learning. However, your model performs poorly on both training and testing datasets. What is the main problem that cause this situation? How to correct the problem? List 2 potential solutions.

Q8. We have seen several types of linear classifiers, such as Logistic Regression, SVM, and Naïve Bayes with Gaussian CCDs with the same variance, which can all be expressed as learning the same function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. How are these linear classifiers different? Which linear classifier is best in terms of accuracy?

Q9. Suppose you are a global market analyst who needs to analyse a product line (a group of similar products with different prices) and get some insight. You want to classify your 50,000 clients according to the product they brought. You have collected 25 features on your clients, e.g. age, income, educational level, frequency of using the product, their scores on other similar products. Among Logistic Regression, kernel SVM and random forest, which classifier will you prefer to use? List 2 reasons for using your selected classifier, and 1 reason for each of the other unselected classifiers.

Q10. You are working for a hospital to build a classifier as a cost-effective screening test for a lung disease based on the patient's demographic data and saliva sample. If the classifier predicts positive, then the patient will do a follow-up CT scan that is more costly and very accurate. You have collected a dataset of 1000 patients, of which 50 have the lung disease. What are the key issues to consider when training and testing your classifier? How do you address these key issues during training and testing?

Q11. The goal of supervised training is to find the weights that best fit the training data (X, Y), To determine what we mean by best fit, we introduce a loss function L and the goal is to find the weights that minimize:

$$\hat{w} = \underset{w}{\operatorname{argmin}}\, L(X, Y) + \alpha\|w\|^2.$$

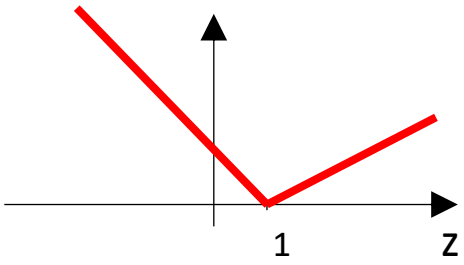
a) What is the purpose of the term $\|w\|^2$?
b) How is α determined?
c) Suppose we replace $\|w\|^2$ with $\|w\|_1$. What will be the effect when training the classifier?

Q12. Consider a linear regression function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+b$, which is trained on the data $(\mathbf{x}_i, y_i)$, $i=1\ldots N$, using the following optimization problem, where the absolute prediction error is the loss function:

$$(\hat{\mathbf{w}}, \hat{b}) = \underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{N} |f(\mathbf{x}_i) - y_i|$$

What will be the effect of using this loss during training? Why?

Q13. Consider the following classification loss function. Describe the properties of the classifier learned with this loss function. Will this be a good classifier? Why or why not? Note: z = y*f(x).



--- END ---