CITY UNIVERSITY OF HONG KONG

Course code & title	:	CS5489 Machine Learning: Algorithms & Applications
Session	:	Midterm, Semester B 2020
Time allowed	:	Two hours (Feb 23, 7:00pm-9:00pm)

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

---

1. This paper consists of 13 questions.

2. Answer ALL questions in the accompanying **"CS5489-midterm-2021-answersheet.docx"**.

---

*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).

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.*

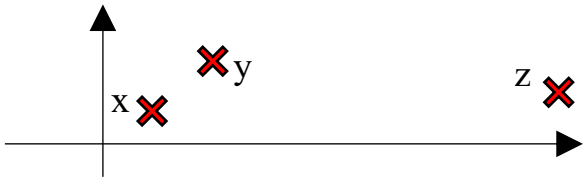
Q1. Which statements are correct? (select all that apply)
A) One SVM model can be used for a three-class classification problem.
B) Logistic regression only forms a linear decision surface.
C) Normalization of features must be done before training a Logistic Regression.
D) Naive Bayes is not appropriate for continuous valued variables.
E) Lasso Regularization can be used for variable selection in Linear Regression.

Q2. Suppose you are working on a classification problem. Your classifier gives low training error but high testing error. The reasons could be: (select all that apply)
A) The training set is too small.
B) The testing set is too large.
C) The distributions of the training set and testing set are different.
D) The classifier is too strong and overfits on the training set.
E) The classifier is too weak to fit on the testing set.

Q3. Which statements about support vector machines are NOT correct? (select all that apply)
A) Kernel SVM cannot handle linearly separable data.
B) Data points lying around the regression tube boundary in support vector regression are called support vectors.
C) Kernel SVM can be interpreted as learning a linear classifier in a high-dimensional space.
D) The kernel function is limited to vector data only.
E) Because SVM only needs support vectors to estimate model parameters, it has good scalability to large datasets.

Q4. What are the benefits of adding regularization terms (L1 or L2) to a linear regression model? (select all that apply)
A) Both L1 and L2 can make the regression model easier to optimize.
B) Both L1 and L2 can make the matrix inversion well-conditioned.
C) Both L1 and L2 make the regression model more robust to outliers.
D) Adding L1 or L2 term can encourage some weights to be close to zero so as to select features.
E) Both L1 and L2 can penalize the training loss of the regression model and prevent overfitting.

Q5. Suppose we have the points x, y, and z below. We compute the Gaussian RBF kernel between the points, k(x,y) and k(x,z), using inverse bandwidth γ. Which statements are correct? (select all that apply)



A) For some choice of γ, both k(x,y) and k(x,z) will be close to 1.
B) For some choice of γ, both k(x,y) and k(x,z) will be close to 0.
C) For any choice of $\gamma > 0$, $k(x,y) \geq k(x,z)$.
D) If k(x,y) is close to 1, then k(x,z) is also close to 1.
E) If k(x,y) is close to 0, then k(x,z) is also close to 0.

Q6. Which statements are true about Bayesian classifiers? (select all that apply)
A) Naïve Bayes classifiers can only model linear decision surfaces.
B) Bayesian classifiers explicitly define the posterior p(y|x).
C) Bayesian classifiers cannot overfit the training data.
D) Bayesian classifiers minimize the probability of making a prediction error.
E) In Bayesian classifiers, the class probability p(y) does not affect the classifier.

Discussion Questions (70 marks)
10 marks each question.

Q7. Consider the classification problem of predicting a person's gender from an image of their face. Discuss the differences between adopting a discriminative classifier or a generative model for classification? How would you interpret the learned generative classifier and learned discriminative classifier?

Q8. Consider the following statement: *If we train a Bayes classifier using infinite training data that satisfies all of the modeling assumptions (e.g., data is from the class-conditional densities, and sampled independently), then it will achieve zero training error over these training examples.*
Is this statement true or false? Give your reasons.

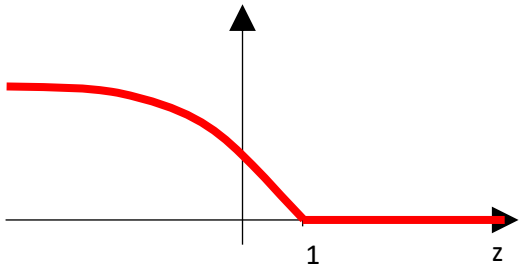
Q9. You have been asked to build a classifier for detecting junk email. You are given a training dataset, which has more junk email than regular email. You trained the model with logistic regression. However, your classification model predicts most emails as spam, so that you will hardly receive any email at all. What are two approaches to fix this problem? Give reasons why these approaches will work.

Q10. Geographical origins of music are related to the audio features in music. Suppose we have 10,597 music segments from different regions of the world, where each audio segment is represented by a 68-dimensional vector, and the region is denoted by latitude and longitude coordinates. We know there are some outliers in the dataset and some regions have apparently more music data than others. Suppose we want to choose Gaussian process regression or random forest regression to predict the region coordinates from the audio feature vector. For this specific regression task, what are the advantages and disadvantages of using these Gaussian process regression or random forest regression?

Q11. Your friend wants to predict the number of taxis used in HK in one hour from various features, such as weather, time of day, day of year, etc. Your friend uses linear regression to fit the data, but finds that the mean squared error (MSE) on the validation set is about the same as a "dummy" regressor that just predicts the mean of the training data. What two pieces of advice would you give to your friend to improve their regressor? Explain why.

Q12. You are working for Tesla on their self-driving car software for their new model. Your goal is to classify objects from the radar scanner. After extensive testing, you find that AdaBoost and Random Forest classifiers both perform equally well at the task. They also both use equal amount of computation (flops). The Random Forest uses 10 trees with depth 10, while Adaboost has 100 weak learners. The new Tesla model has a special 100-core processor for running the self-driving car software. Which classifier do you recommend for implementation in the new car? Explain why.

Q13. Consider the following classification loss function. Describe the properties of the classifier learned with this loss function. Note: z = y*f(x).



--- END ---