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

---

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).

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: Xiaoyuan Zhang
EID: xzhang2523
Student ID: 56287671

**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:** ACD

**Q2 ANSWER:** CE

**Q3 ANSWER:** B

**Q4. ANSWER:** AC

**Q5. ANSWER:** BD

**Q6. ANSWER:** CDE

---

**Discussion Questions (70 marks)**
10 marks each question.

---

**Q7 ANSWER:**

(1) **Problems:**The training data is not fit very well because the model may be too simple or too weak.

**Solution:** We can introduce non-linear components to our model. For example, we can add a non-linear kernel or a polynomial mapping for the features. We can try more powerful models. We can try Adaboost, RandomForeset, XGboost, Gaussian Process regressor.

(2) **Problems:** The dimension of the feature may be low. There may be not enough features.

**Solution**: We add some more feature dimensions. We can collect more useful features.

(3) **Problems:**The hyper-parameter in the model, like learning rate, regularization weight may be not fine-tuned.
**Solution**: Use various way including cross validation to fine tune the best hyper-parameter.

**Q8 ANSWER:**

1 SVM is different with Logistic Regression (LR): (1) SVM is to maximize the decision margin and LR is to push the decision boundary away from all the training data.
2 Naïve Bayes with Gaussian CCDs with the same variance (NB) is different with LR: the decision function of NB is also depended on the prior p(y). LR may use other ways to do deal with the prior. But they are still different.
3 NB is a generative model and SVM, LR are discriminative models.
4 Because of the **No-free lunch theorem**, there is no a best classifier. The accuracy is also depended on the data.

**Q9 ANSWER:**

<ANSWER HERE>

1. Unselected classifiers:

1.1 stic Regression (LR) is not selected because LR is a linear model. There are 25 features. So it is very hard to separate the data just by a linear decision boundary.
1.2 SVM is not selected because an intermediate step in SVM is to build an 50000x50000 matrix. It is too expensive.

2. We select the Random Forest classifier. The reasons are:

a) The Random Forest classifier is very suitable for **parallel computing**. Since we have 50000 training dataset, it is very good if we can do **parallel computing.**
b) Since Random Forest usually use the voting method to predict the final result. It is less likely to **overfit** the data.
c) The prediction of Random Forest is also very fast comparing with SVM.
d) Random Forest can determine the importance of a feature. RF even can give some interpretation across different features. Hence, RF is more suitable for interpretation.
e) RF is very easy for implementation. It usually needs less coding.

**Q10 ANSWER:**

<ANSWER HERE>

The key issue is **Unbalance**. There are both data unbalance and classifier unbalance in this problem.

**Solutions**:

1. We should add weights for **training**. Since we have only 50 patients with a lung disease. We should use a larger weighting factor for patients with a lung disease.
2. For **testing**, usually the threshold for predicting a patient having a lung disease is T>0. We lower this threshold. For example, we can predict a patient having a lung disease by T>-0.2. (We assume the label for a lung disease is +1, and having not a disease is -1).
3. We can use **data augmentation** during training. For example, we can randomly choose a large part features of the 50 patients with a lung disease. Then we can make more samples of patients with a lung disease. The data augmentation skills will also help with the unbalance classification problems.

---

**Q11 ANSWER:**

<ANSWER HERE>

(a) The purpose of the term $||w||^2$ is to prevent overfitting. The term $||w||^2$ will make some $w$ close to zero, hence reduce some kind of complexity of the model.

(b) $\alpha$ can be determined by cross validation. We can use grid searching to select the best $\alpha$.

(c) (1) The training may be more difficult because $||w||$ is not differentiable and $||w||^2$ is differentiable. (2) The term $||w||$ will make some $w$ **exactly** be zero and the term $||w||^2$ will make some $w$ close to zero.

---

**Q12 ANSWER:**

<ANSWER HERE>

(1) This function is not differentiable. So there may be some unstablily during the training.

(2) This function is more robust for the outliers comparing with $||f(x_i) - y_i||^2$. Because if $|f(x_k) - y_k|$ is very big (maybe an outlier), the function $|f(x_i) - y_i|$ only punish $|f(x_k) - y_k|$ in the order of 1. There is no extra focusing punish the very large errors comparing with $||f(x_i) - y_i||^2$.

(3) There is no a regularization term for the weigh w. It means the total loss function can be trained to be very small value finally. However, it may not behave well on the testing dataset.

---

**Q13 ANSWER:**

(1) The classifier only can predict $f(x) = \pm 1$ only since there is only one minimal value at $z = 1$.

(2) It is more robust to outliers comparing with a Logistic regressor. The punish the outliers with a linear loss. The loss for outliers and other misclassified data are a linear function with $z$.

(3) It will also punish the very well classified data ($z > 1$) back to $z = 1$. It means this classier does not encourage the very well classified data. It does not make sense in reality.

It is **NOT** a good classifier.

---

--- END ---