IMPORTATNT INSTRUCTION: Please coply the follwing sentences to your answer sheet and sign your name.

"*I certify that I take the honor code of CUHK seriously, and I take this final examination without asking another person or searching for related answer in the Internet.*"

2 Questions

1. (**Multiple choice**) Which of the following problem is *not* supervised learning:

   (a) Predict an article as to whether it belongs to *sport*, *entertainment* , *editorial* or *others*.

   (b) Determine whether a customer is a *high risk*, *moderate risk* or *low risk* customer who wants to apply for a loan from a bank.

   (c) Determine the cluster of a cancerous gene among K different clusters of genes.

   (d) Determine whether a user will click on a given advertisement link

   (e) Determine whether an image belongs to a student in *computer science*, *computer engineering*, *information engineering* and others.

2. (**Multiple choice**) In class, we discussed PCA and feature embedding method. If the number of sample points $N = 2000$ and the number of independent features is $d = 2500$, and that I know features are all independent. Which method should I use so that I can have a good balance of accuracy and low computational complexity?

   (a) Feature embedding because $N < d$

   (b) PCA because I only understand PCA

   (c) Does not matter, I can use either PCA or feature embedding

   (d) Reduce number of features d to 2000, then use feature embedding

   (e) Expand number of samples N to 3000, then use PCA.

3. (**Multiple choice**) When we do linear regression (i.e., least square method), we want to add the regularization component because:

   (a) Reduce computational complexity

   (b) Improve accuracy

   (c) Avoid over-fitting

   (d) Avoid under-fitting

   (e) Compensate in case there are not sufficient data

4. (**Multiple choice**) In class, we discussed decision tree. In this method, we need to find a way to partition the sample points within a decision node into different disjoint partitions. In the lecture, we suggested to use entropy function to determine the threshold of a single feature so to so the partition. Assume we do not use entropy function and our input has $d > 10$ features, which of the following function is appropriate for us to use in decision tree:

   (a) A multi-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$ wherein f is continuous and increasing

   (b) A multi-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$ wherein f is continuous and decreasing

   (c) A multi-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$ wherein f is continuous and convex.

(d) A multi-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$ wherein f is continuous and concave.

(e) A piece-wise linear multi-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$.

5. (**Multiple Choice**) In a multi-class logistic classification where $K > 2$, we use the soft-max technique and the log likelihood cost function. The reasons for using these techniques are:

   (a) Estimate the mutual information of sample points in different classes and then apply gradient method

   (b) Approximate the class probability of sample points so to do the one-round training

   (c) Use the soft-max as an estimate of class probability and apply it to a negative convex function such that we can apply gradient method.

   (d) We need soft-max for an iterative procedure for expectation maximization

   (e) None of the above

6. (**Multiple Choice**) You are given the following figure wherein we have two classes of training points (they are in RED DIAMOND and BLUE CIRCLE). Which classification method listed below will provide the given decision boundary?

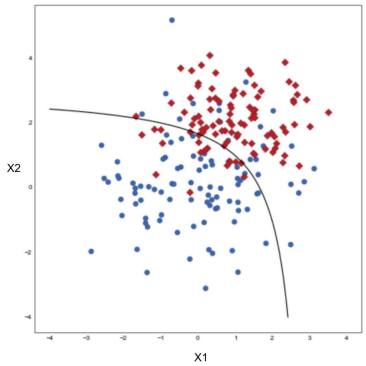

Figure 1: Figure 1

   (a) Logistic classification with $K = 2$

   (b) Linear Discriminant Method

   (c) KNN with $K = 1$.

   (d) Quadratic Discriminant Method

   (e) PCA

7. (**True or False**) In class, we said that classification is a *discrete version* of regression, therefore, classification machine learning is just a special case of regression.

   (a) True

   (b) False

8. (**True of False**). Assume the feature space is three dimensional (of $d = 3$), John claims that we can always find a set of four points in $I\!R^3$ that can be shattered by a 2D hyperplane .

   (a) True

   (b) False

9. (**True or False**) You are given a data set $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. The mean is $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 10$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = 8$. When applying the linear regression method, you find that the slope of the least square is 1.5. John said that the intercept should be greater than 23, true or false?

   (a) True

   (b) False

10. (**True of False**) For the sample points below wherein we have four points with $K = 2$ classes We can use either 1NN classifier or linear discriminant to achieve zero training error.

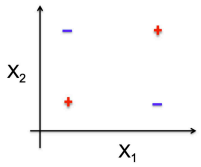

Figure 2: Figure 2

   (a) True

   (b) False

11. (**True of False**) Assume our inputs have seven features and we have six data points. Our training set is
    $X = \{\ (5, 5, 0, 0, 5, 5, 0),\ (0, 0, 1, 2, 0, 0, 1),\ (0, 0, 5, 10, 0, 0, 5),\ (0.2, 0.2, 0, 0, 0.2, 0.2, 0),$
    $(-2, -2, 0, 0, -2, -2, 0),\ (0, 0, 0.25, 0.5, 0, 0, 0.25)\}$. Let say we want to use PCA to reduce its dimension. Then the final dimension size (k) should be $k = 6$ if we want to make sure that PoV is at least greater than 0.8 because $6/7 > 0.85$.

   (a) True

   (b) False

12. (**Essay question**) In class, we discussed the logistic classification and how to use the gradient decent method to find the weighting of features. Assume that the input has d features and there are only $K = 2$ classes. We want to re-drive the *gradient decent update rules* when we apply *regularization*, i.e., we use the L_2 norm or penalty as regularization.

    (a) State the log likelihood function.

    (b) State the error function we wish to reduce.

    (c) Derive the gradient decent update rules for weight w_i, where $i = 0, 1, \ldots, d$, wherein weight w_i indicates the impact of feature x_i in the classification process.

13. (**Essay question**) Assume that our samples are high dimensional points (i.e., d is large) and we use PCA to reduce it to $k = 10$ dimensions. After this step, we found that all the 10 new dimensions have continuous values (e.g., in other words, each feature in the transformed dimension is *not* from discrete domain, but rather, continuous domain). Describe in detail, how we can now use parametric method to train our model to do classification. In particular, discuss how we can compute the correlation matrix estimation, and when a new point arrive, what procedure we need to do so to make a classification prediction (assume in general, we have $K > 2$ classes).

14. (**Essay question**) You are given five *loaded* dices, D_i, $i = 1, 2, 3, 4, 5$, wherein each dice can have six possible outcomes (or 1, 2, 3, 4, 5, or 6). Since these dices can be loaded, the probability of seeing one of the possible outcomes may *not* be $1/6$.

    To play this game, you can request for a dice, and the system will randomly pick one dice out of the five dice for you. Note that you *DO NOT* know which dice was selected. Then the system will toss this dice $L > 0$ times and inform you these L outcomes. You can play this game as many times as you want.

    Device a procedure to estimate the probability of dice D_i having output k, or $P(k|D_i)$ where $k = \{1, 2, 3, 4, 5, 6\}$ and $i = 1, 2, 3, 4, 5$.

-End-

Wishing you all a wonderful summer. Do remember to read some good books and learn some new knowledge.