光华管理学院 Guanghua School of Management | UG&PG 本科研究生项目

机器学习与人工智能
Machine Learning and Artificial Intelligence

2021 Fall

Final Exam

姓名：_____

学号：_____

Important Notes

1. Write down your full name and ID first (see above)
2. You can answer the questions in either **Chinese** or **English**
3. Individual-effort, closed-book, closed-notes. A calculator is allowed.
4. 120 minutes.
5. Components of this exam: (total points are 100 + 5 bonus)
   - 8 true/false questions (2.5 points each). (20 points in total)
   - 5 multiple choice questions (4 points each). Pick the one that you think is correct. (20 points in total)
   - Short answer questions (60 points in total). Provide clear responses with justifications when necessary. It is not **how much** you write but **what** you write that matters. So be precise and concise in your responses. Readability/neatness of your writing will be considered.
   - One bonus question (5 point).
6. Use your time wisely -- if got stuck on a question, move on and come back to it later.

GOOD LUCK!

QI: Determine whether each of the following statements is true or false. If false, then provide the right statement instead. (2.5' each, 20 points)

1. Changing Sigmoid activation to ReLu will help to get over the vanishing gradient issue.

2. In boosting, we train multiple weak learners and use the best one that achieves the highest performance on the validation dataset as the final classifier at test time.

3. SVM with no kernel cannot be applied when the data are not linearly separable.

4. One disadvantage of Q-learning is that it can only be used when the learner has prior knowledge of how its actions affect its environment.

5. Overfitting is less likely to happen when the size of the feature space is much larger than the number of training examples.

6. The Gradient Descent technique can always find an optimal solution.

7. In general, k-NN performs better in a very high dimensional feature space than in a low dimensional feature space.

8. When growing a Decision Tree, the splitting at each step is done by choosing the variable that gives the highest entropy.

QII: Multiple Choice Questions (4' each, 20 points)

1. For a polynomial regression, which one of these structural assumptions is the one that most affects the trade-off between underfitting and overfitting?

    A. Whether we learn the weights by matrix or gradient descent
    B. Whether the outcome is numerical or categorical
    C. The polynomial degree
    D. The use of a constant-term unit input

2. Which of the following is not the reason for Deep Learning recently taking off?

    A. Neural networks are a brand-new field
    B. We have access to more computational power
    C. We have access to more data
    D. More optimization algorithms have been introduced

3. Which of the following is true for reinforcement learning?

    A. automated vehicle is an example of reinforcement learning
    B. exploitation strategy allows us to obtain more information of environment
    C. Q-learning requires the full knowledge of the environment
    D. policy should be deterministic given a state value

4. Which of the following techniques can NOT reduce/avoid overfitting in a Decision Tree model?

    A. choose the largest information gain to split a node
    B. prune the tree
    C. reduce the depth of the tree based on validation data
    D. stop growing tree when information gain is less than a threshold

5. Which binary classifier is able to correctly separate the training data given in Figure 1?

   A. Logistic regression
   B. Linear SVM
   C. Decision Tree
   D. 3-NN classifier (with Euclidean distance)

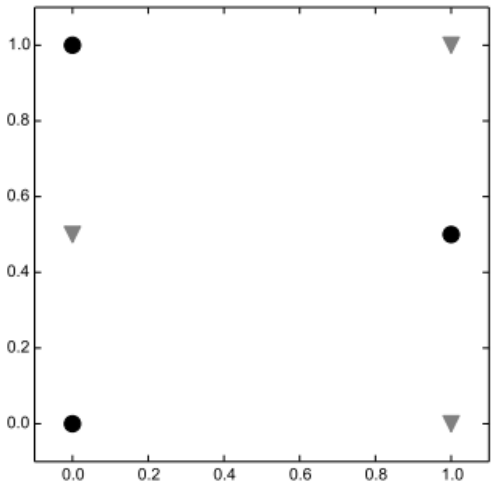


Figure 1

Q III: Short Answer Questions (60 points)

Part A. (17 points)

1. Distance is a key notion underlying many machine learning algorithms, such as k-nearest neighbor (k-NN). What problem is there with comparing consumers using regular Euclidean distance, for example when they are described by age (in years), income (in RMB), and number of credit cards? How can this problem be fixed? (5 points)

2. ROC curves can be more effective for assessing model quality than the percent of classifications that are correct. Give two different reasons why. (6 points)

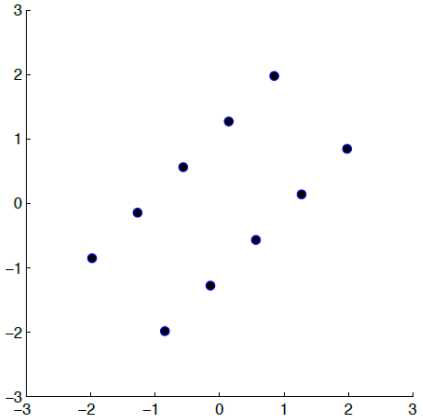
3. A student is working on a machine learning project about spam detection. The data includes 200 labeled emails, 90% of which are used for training and 10% for validating the model. The student experiments with over 100 different learning models, training each one on the training set and recording the accuracy on the validation set. The student's best model achieves 90% accuracy on the validation set. Would you pick the student's solution for protecting a corporate network from spam? Give at least two reasons for your choice. (6 points)

Part B. PCA (16 points)

PCA should be used with caution for classification problems, because this unsupervised learning technique does not take information about classes into account. In this problem, you will show that, depending on the dataset, the results may be very different.

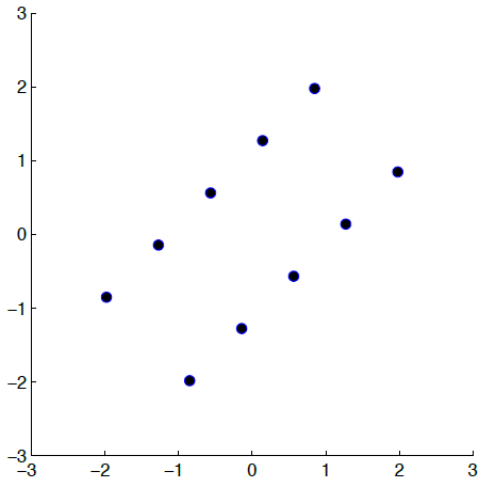
Suppose that the classification algorithm is 1-nearest-neighbor, the source data is 2-dimensional and PCA is used to reduce the dimensionality of data to 1 dimension. There are 2 classes (+ and -). The data points (without class labels) is pictured on the plots below.

(1) On the data plot, draw the direction that PCA will project the data points to (i.e., the first principal component) (6 points)

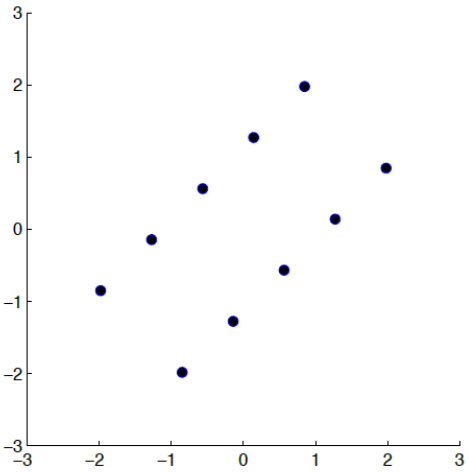


(2) For each of the plots (the plots are identical), provide the labeling for all the 10 data points so that 1-NN will have the following training accuracy: (10 points)

a) Original 2D data: 0% accuracy; 1D projected data from PCA: 100% accuracy
b) Original 2D data: 100% accuracy; 1D projected data from PCA: 0% accuracy



a)



b)

Part C. Regression (12 points)

You are asked to use a regularized linear regression to predict the target variable $Y \in R$ from the eight-dimensional feature vector $\boldsymbol{X}$. You define the model $Y = \boldsymbol{\omega}^T \boldsymbol{X}$, where $\boldsymbol{\omega}$ is the coefficient vector. In total, you have n training data points, each of which is denoted as i. Then you recall from class the following three objective functions to estimate $\boldsymbol{\omega}$:

$$\min_{\omega} \sum_{i=1}^{n} (y_i - \omega^T x_i)^2 \quad \text{(A.1)}$$

$$\min_{\omega} \sum_{i=1}^{n} (y_i - \omega^T x_i)^2 + \lambda \sum_{j=1}^{8} \omega_j^2 \quad \text{(A.2)}$$

$$\min_{\omega} \sum_{i=1}^{n} (y_i - \omega^T x_i)^2 + \lambda \sum_{j=1}^{8} |\omega_j| \quad \text{(A.3)}$$

Based on the above settings, answer the following two questions:

1. For large values of λ in (A.2), will the model complexity decrease or increase?

2. The following table contains the estimated coefficients learned for the three objective functions.

	Column A	Column B	Column C
ω_1	0.38	0.50	1.60
ω_2	0.23	0.20	1.30
ω_3	-0.02	0.00	-1.10
ω_4	0.15	0.09	1.20
ω_5	0.21	0.00	1.30
ω_6	0.03	0.00	1.20
ω_7	0.04	0.00	1.02
ω_8	0.12	0.05	1.26

Beside each function, write the most appropriate column label (A, B, or C):

Objective function (A.1):
Objective function (A.2):
Objective function (A.3):

Part D. Clustering (15 points)

(a) We would like to cluster the points in Figures 2 using k-means. We set k = 2. We perform several random restarts for each algorithm and chose the best one as discussed in class. Please show the resulting cluster centers in the figure.

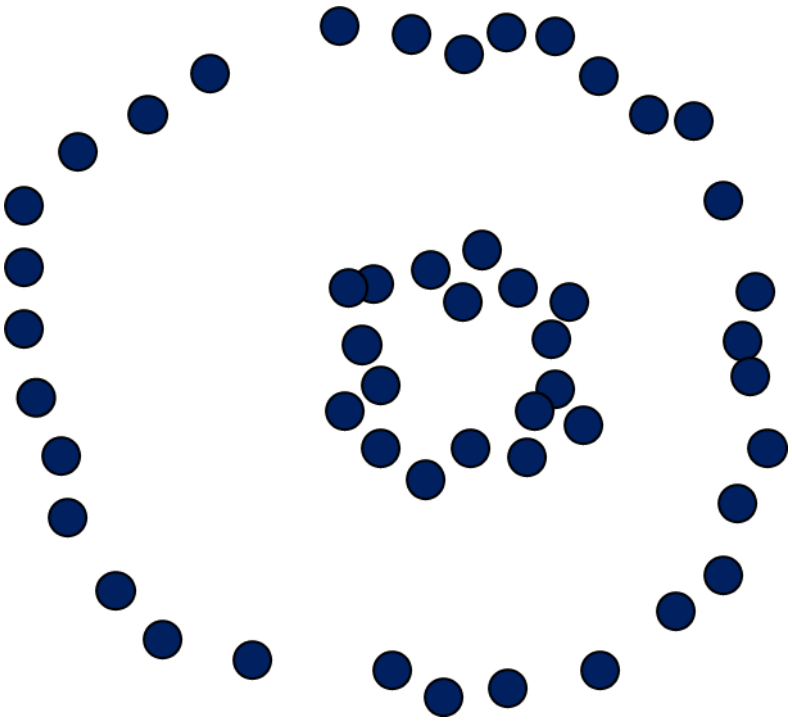


Figure 2 K-means

(b) Next we would like to use the hierarchical clustering on the same dataset. We will use the Euclidian distance as the distance function. In both cases we cut the tree at the second level to obtain two clusters. For two of the linkage models learned in class, single (i.e., MIN distance) and group average link, circle the resulting groups of points on each of the figures (Figure 3 - single link, Figure 4 - average link).

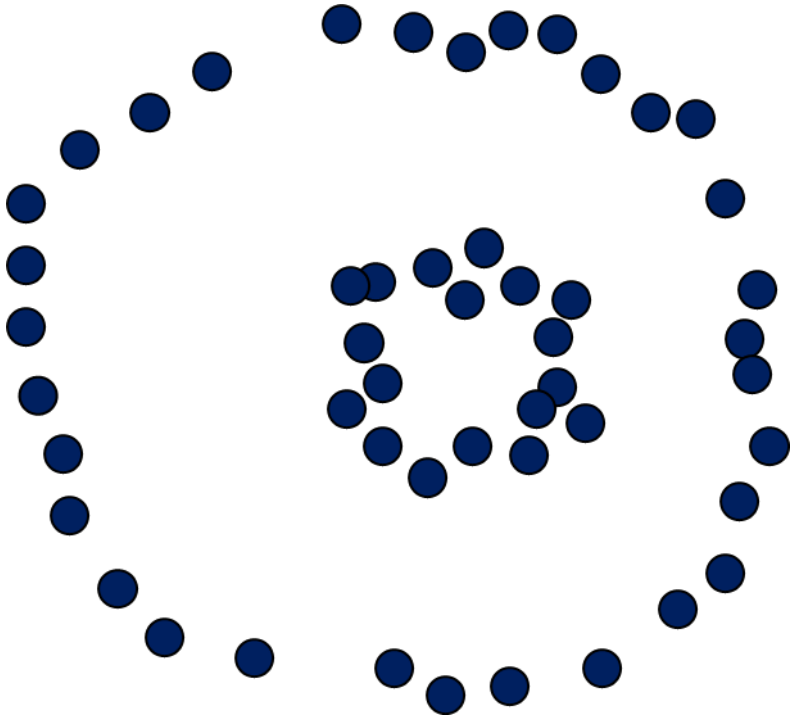


Figure 3 Single link

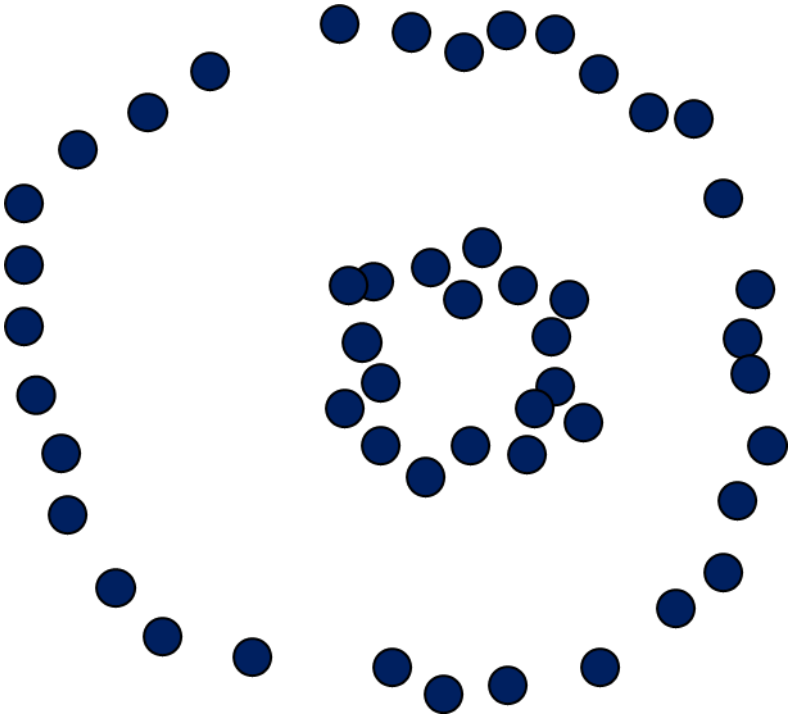


Figure 4 Average link

Bonus Question: Reinforcement Learning (MDP) (5 points)

Consider a robot that is moving in an environment. The goal of the robot is to move from an initial point to a destination point as fast as possible. However, the robot has the limitation that if it moves fast, its engine can overheat and stop the robot from moving. The robot can move with two different speeds: slow and fast. If it moves fast, it gets a reward of 10; if it moves slowly, it gets a reward of 4. We can model this problem as an MDP by having three states: cool, warm, and off. The transitions are shown in below. Assume that the discount factor is 0.99 and also assume that when we reach the state off, we remain there without getting any reward.

s	a	s'	P(s' | a,s)
Cool	Slow	Cool	1
Cool	Fast	Cool	1/2
Cool	Fast	Warm	1/2
Warm	Slow	Cool	1/2
Warm	Slow	Warm	1/2
Warm	Fast	Warm	1/2
Warm	Fast	Off	1/2

1. Consider a policy when the robot always moves slowly. What is the value of $V(cool)$ under this conservative policy?

Hint: Use the value function form to derive the discounted sum of rewards $V(cool)$. Then solve the equation to get the numerical value. (2 points)

2. What is the optimal policy for each state? Justify your answer. (3 points)

Hint: It is not necessary to use math to solve this problem. You can explore the answer based on the probability, the risk of turning off, and the discount factor.