Student ID:_____ **Student Full Name:** **(CN)** _____ **(EN pinyin)** _____

Instructions:

- Provide your students information in the fields above for the quiz to be graded
- Your **hand-writing must be legible** for grading purposes.
- Consulting with classmates is **forbidden**
- There are a total of 3 short exercises and 7 multiple choice questions.

Question 1 (17 points): Explain what are supervised learning and unsupervised learning and what are their main differences. (Answer in about 5 lines)

Supervised learning: The learner is provided with a set of data inputs together with the corresponding desired outputs. Classification and regression are supervised learning

Unsupervised learning: Training examples as input patterns, with no associated output. Clustering is unsupervised learning.

Main differences: unsupervised learning has no "teacher", supervised learning uses labeled input and output data, while an unsupervised learning does not.

Question 2 (17 points): Explain the differences between classification and regression and provide **1 example for each** (you can use figures to explain their differences). (Answer in about 5 lines)
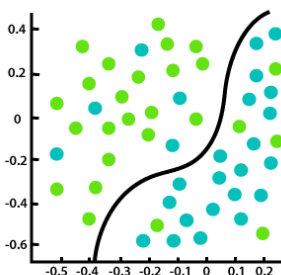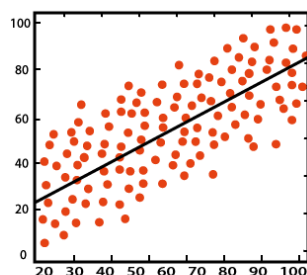
Classification   Regression

In supervisor learning, when the target variable that we're trying to predict is continuous, we call the learning problem a regression problem. When the target variable can take on only a small number of discrete values, we call it a classification problem. The prediction of iris class is a classification problem. The prediction of next week's temperature is a regression problem.

Question 3 (17 points): Explain the meaning and the components of the loss function
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$
Why does it have a power of 2 expression? (Answer in about 4 lines)

MSE stands for Mean squared error, it is the average of the squared difference between the actual y_i and predicted $f(x_i)$ values. It has a power of 2 expression in order to consider both under and over prediction as error. Without power 2 expression, the sum of over and under predictions will cancel out each other so it is not a measure of performance.

Multiple Choices (Only one choice is correct. Each question is worth **7 points**). Report answers in the table:

Question	1	2	3	4	5	6	7
Answer	B	C	B	C	A	D	D

1. Which of the following can be considered unsupervised learning?
   - A. Sale Forecast
   - B. Find communities of internet users
   - C. Predict spending on food
   - D. Determine whether market sentiment is good or bad
   - E. House price estimation

2. Which one of the following is an application of "regression" (other than "classification")?
   - A. Determine whether market sentiment is good or bad
   - B. Determine whether there is a person in a picture
   - C. Predict the future price of a stock
   - D. Predict whether a person is married
   - E. Predict whether is will rain

3. We should measure the accuracy of a classifier based on …
   - A. the training set
   - B. the test set
   - C. the whole data set
   - D. None of them

4. In a regression we can use the following measure to assess prediction performance?
   - A. Gini Index
   - B. Manhattan Distance
   - C. R-Squared
   - D. Correlation

5. R-Squared is a value is always:
   - A. between 0 and 1
   - B. bigger than 0.5
   - C. smaller than 0.95
   - D. between -1 and 1

6. Given the predicted weather results as follows, for which record the prediction label is **wrong**?

Record	Predicted Rainy Day	Actual Weather (Rainy or Not)	Prediction label
A.	Yes	No	FP
B.	No	No	TN
C.	Yes	Yes	TP
D.	No	Yes	FP

7. In KNIME, the node "Partitioning" perform the following on the data: …
   - A. Bootstrapping
   - B. Calculate the accuracy of the model
   - C. Cross-validation
   - D. Divide the data for training and testing purpose