True/False Questions

1. **TF2.1** If we would like to use imputation for data cleaning, we may consider replacing the missing value using the average from other samples.
   a) True
   b) False

Ans: True

2. **TF2.2** Data integrity concerns the maintenance and the assurance of data accuracy and consistency.
   a) True
   b) False

Ans: True

3. **TF3.1** Assume we have two urns. The first contains 6 white and 5 black balls, and the second contains 4 white and 5 black balls. We flip a fair coin, and pick the first urn if the coin shows a head while we pick the second urn if the coin shows a tail. We then draw a ball from the chosen urn. The probability of eventually getting a white ball is greater than a black ball.
   a) True
   b) False

Ans: False. Probability of drawing a white ball is 1/2*6/11+1/2*4/9, which is smaller than the one of a black ball which is 1/2*5/11+1/2*5/9.

4. **TF6.1** The One Hot Encoding (OHE) can be applied to a binary target.
   a) True
   b) False

Ans: a)

5. **TF6.2** The polynomial model can approximate any continuous real-valued function on a closed and bounded interval to any degree of accuracy.
   a) True
   b) False

Ans: a)

6. **TF6.3** There are three samples of two-dimensional data points $\mathbf{X} = \begin{bmatrix} 1 & 3 \\ 4 & -1 \\ 3 & 3 \end{bmatrix}$ with the corresponding single output target vector $\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$. Suppose you want to use a full third-order polynomial model to fit these data, the polynomials model has 9 parameters to learn.
   a) True

b) False

Ans: b) The model has 10 parameters to learn (3rd order polynomial with two-dimensional data): offset, $x_1, x_2, x_1^2, x_2^2, x_1x_2, x_1^2x_2, x_2^2x_1, x_1^3, x_2^3$

7. **TF7.1** Suppose the model parameters are given by $\boldsymbol{w} = [w_1, \dots, w_d]^T$. Then an inclusion of the L2-regularization during training reduces the value of $\sum_{i=1}^{d} w_i^6$.
   a) True
   b) False

Ans: False

8. **TF8.1** Suppose we are minimizing a cost function C(w) with respect to w using a gradient descent algorithm. We observe the following: At iteration 1, C(w) is -15. At iteration 2, C(w) becomes 15. At iteration 3, C(w) becomes -7. At iteration 4, C(w) becomes 12. At iteration 5, C(w) becomes -13. To improve our algorithm, we should consider decreasing the learning rate.
   a) True
   b) False

Ans: True. C(x) seems to be oscillating. We should consider decreasing the step size.

9. **TF9.1** Compared with a decision tree, a random forest has lower squared bias and the same variance
   a) True
   b) False

Ans: False. Random forest has lower variance, but squared bias stays the same

10. **TF10.1** To enhance the model generalization capability, we should wisely select samples from the test set and utilize them in the training stage, which will potentially improve the performances.
    a) True
    b) False

Ans: False. We should never touch test set during training.

11. **TF10.2** A higher training accuracy not necessarily leads to a higher test accuracy; however, in most cases, a higher training accuracy indeed leads to a higher validation accuracy.
    a) True
    b) False

Ans: False. There is no connection between higher training accuracy and higher validation accuracy.

12. **TF10.3** Suppose we would like to develop a human face detector, where the positive class corresponds to face and negative class corresponds to non-face. In this case, *false positive* of the positive class denotes the face(s) missed by the detector.
    a) True
    b) False

Ans: False. False negative denotes the face(s) missed by the detector.

13. **TF10.4** We have built a binary classifier. During testing, it is possible that the sum of the false-positive-rate and the false-negative-rate is greater than 100%.
    a) True
    b) False

Ans: True.

14. **TF11.1** In the naïve K-means method, the optimal K is learned automatically.
    a) True
    b) False

Ans: False. K is set by hand in naïve K-means.

15. **TF11.2** Confusion matrix is a popular metric for evaluating regression tasks.
    a) True
    b) False

Ans: False. Confusion matrix is a popular metric for evaluating classification tasks.

16. **TF12.1** Recent research in neural networks mainly focus on manually setting the parameters of the networks smartly, so that they can generalize well to unseen data.
    a) True
    b) False

Ans: False. Parameters of the networks are learned by backpropagation.

Multiple Option Questions (1.5 marks * 16 Questions)

1. **MCQ1.1** Which of the following statement(s) is/are true about machine learning? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
    a) Jack went hiking with his friends and collected a bag of colorful stones. He would like to group these stones by color. As such, Jack is doing a supervised classification on the stones.

b) Jack went hiking with his friends and took 100 pictures of birds. One of his friends David, who is a scientist working on bird species, looked at all the pictures and told Jack there are in total five bird species. He picked one picture from each species, showed it to Jack, and told Jack the name of the species. Jack went home and categorized the rest of the pictures based on Jack's annotation. As such, Jack is doing a supervised classification on the bird pictures.
c) Jack went hiking with his friends and collected a bag of colorful stones. After he went home, he measured the weight, size, and color of each stone, before grouping them into different categories. The weight, size, and color measurement step can be thought as the feature extraction step.
d) None of the others.

Ans: b) c)

2. **MCQ2.1** Which of the following statement(s) is/are true about data wrangling? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
   a) Data wrangling concerns transforming "raw" data into an appropriate form for downstream purposes such as analytics.
   b) Assume we have three classes, *car, truck, bicycle*, and we would like to use one-hot encoding to denote the three classes. One possible representation is as follows: 'car'=[0,0,1], 'truck'=[1,0,0], 'bicycle'=[1,1,1].
   c) There is no need to conduct data wrangling if we know a priori that, there are no noisy samples in the data records.
   d) None of the others.

Ans: a)

3. **MCQ3.1** The discrete random variable X has the following probability mass function,

X	1	2	3	4
P[X]	0.1	2k	2k	5k

What is value of k?
a) k=0.1
b) k= 0.2
c) k=0.3
d) k=0.02

Ans: a). 0.1+9k=1, therefore k=0.1

4. **MCQ5.1** This question is related to the understanding of linear systems and partial derivatives. Which of the following statements below is/are correct? **There can be more than one answer.**

a) In under-determined linear systems, the number of parameters is greater than the number of unknown equations.

b) The system $\begin{bmatrix} 1 & 4 \\ 2 & 7 \\ -3 & 11 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2.5 \\ 4 \end{bmatrix}$ has no *exact* solution but an approximated solution is available using the left inverse.

c) If $\mathbf{f}(\mathbf{x})$ is a vector-valued function of *size p* x1 and $\mathbf{x}$ is an *m* x1 vector, then differentiation of $\mathbf{f}(\mathbf{x})$ with respect to $\mathbf{x}$ is an *m x p* matrix.

d) Consider the linear system $\mathbf{X}\mathbf{w} = \mathbf{y}$, $\mathbf{X} \in R^{m \times d}$ is the input data matrix, $\mathbf{w} \in R^{d \times 1}$ is the parameter vector, and $\mathbf{y} \in R^{m \times 1}$ is the target vector. If $d > m$, the system has more equations than parameters.

e) None of the others.

Ans: a) & b).

a) True. In under-determined linear systems, the number of equations is greater than the number of unknown parameters.
b) True. It is an over-determined system and left inverse exists.
c) False. If $\mathbf{f}(\mathbf{x})$ is a vector function of *size p* x1 and $\mathbf{x}$ is a *m* x1 vector, then differentiation of $\mathbf{f}(\mathbf{x})$ w.r.t. $\mathbf{x}$ is a *p x m* matrix.
d) False. If $d > m$, there are more parameters than equations.

5. **MCQ6.1** This question is related to the understanding of linear systems, ridge regression, and polynomial regression. Which of the following statements below is/are correct? **There can be more than one answer.**

   a) A set of linear equations is written as $\mathbf{X}\mathbf{w} = \mathbf{y}$ where $\mathbf{X} \in \boldsymbol{\mathcal{R}}^{5\times3}$ is the input data matrix, $\mathbf{w} \in \boldsymbol{\mathcal{R}}^{3\times1}$ is the parameter vector, and $\mathbf{y} \in \boldsymbol{\mathcal{R}}^{5\times1}$is the target vector. There are three simultaneous equations in this set of equations.

   b) The ridge regression can be applied to multi-target regression.

   c) The solution for a ridge regression problem with feature vectors in **X** and targets in **y** can be written as $\widehat{\mathbf{w}} = (\mathbf{X}^\mathbf{T}\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\boldsymbol{y}$ for *λ>0.* As λ increases, $\widehat{\mathbf{w}}^{\boldsymbol{T}}\widehat{\mathbf{w}}$ increases.

   d) Given four samples of two-dimensional data points $\mathbf{X}$ and the corresponding target output **y**, the 2nd order polynomial regression system is an under-determined system.

   e) None of the others.

Ans: b & d).

Reason:

a). There are five equations. False.
b). Yes. True.
c). False. As λ increases, $\widehat{\mathbf{w}}^{\boldsymbol{T}}\widehat{\mathbf{w}}$ decreases.

d). An 2-input 2nd order polynomial regression system has 6 parameters:

$$f_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_{12}x_1x_2 + w_{11}x_1^2 + w_{22}x_2^2.$$

This gives an under-determined system since there are 4 samples here. True.

6. **MCQ6.2** You are given a collection of 5 training data points of three features (x_1, x_2, x_3) and their corresponding class labels which are stacked as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 3 & -2 \\ -4 & 0 & -1 \\ 3 & 1 & 8 \\ 2 & 1 & 6 \\ 8 & 4 & 6 \end{bmatrix}, \mathbf{Y\ class} = \begin{bmatrix} class\ 1 \\ class\ 1 \\ class\ 2 \\ class\ 3 \\ class\ 3 \end{bmatrix}.$$

We would like to predict the class labels of new test samples. Which of the following statements below is/are correct? **There can be more than one answer.**

a) If we would like to perform multi-class classification using the optimal linear regression model with an offset term, the estimated parameter matrix **W** is of dimension 3 x 4.

b) Given a new sample $(x_1, x_2, x_3) = (1, -2, 4)$, the predicted class label using the optimal linear regression model with an offset term is class 3.

c) If we would like to perform multi-class classification using the optimal second-order polynomial model, the corresponding polynomial expansion matrix **P** (with respect to **X**) is of dimension 5 x 10. Therefore, it is an under-determined system.

(d) For two new samples $\mathbf{X}_{new}$, after learning and testing using some polynomial model, the prediction is $\mathbf{Y}_{new} = \mathbf{P}_{new}^T\widehat{\mathbf{W}} = \begin{bmatrix} -0.58 & 0.40 & -0.67 \\ 1.25 & 0.10 & -0.20 \end{bmatrix}$. Therefore, the class label prediction for the first sample is class 3 and the second sample is class 1.

e) None of the others.

Ans: b) & c)

(a) False. Need to add offset to the features so that $\mathbf{X} = \begin{bmatrix} 1 & 1 & 3 & -2 \\ 1 & -4 & 0 & -1 \\ 1 & 3 & 1 & 8 \\ 1 & 2 & 1 & 6 \\ 1 & 8 & 4 & 6 \end{bmatrix}$

The system is an over-determined one,
and hence use $\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} =$

```
[[-0.01005025 -6.03015075  7.04020101]
 [-0.24874372 -1.74623116  1.99497487]
 [ 0.44723618  3.34170854 -3.78894472]
 [ 0.03015075  1.09045226 -1.12060302]]
```

It is a 4x3 matrix. a) is False.
The output prediction is $\mathbf{y}_t = \mathbf{x}_t^T\widehat{\mathbf{w}} =$[-1.03266332 -10.09798995 12.13065327]
Therefore, the category prediction is class-3.

b) True (see above).

(c) True. The polynomial of the 2nd order with three variables has the following expansion terms : $[1 \quad x_1 \quad x_2 \quad x_3 \quad x_1x_2 \quad x_1x_3 \quad x_2x_3 \quad x_1^2 \quad x_2^2 \quad x_3^2 \quad]$
The $\mathbf{P}$ matrix is thus of 5x10 dimension => an under-determined system

$$\widehat{\mathbf{w}} = \mathbf{P}^T(\mathbf{P}\mathbf{P}^T)^{-1}\mathbf{y}$$

d). False. The class label prediction for the first sample is class 2 and the second sample is class 1. The maximum value of each row.

```
#Python code#
import numpy as np
from numpy.linalg import inv
from sklearn.preprocessing import PolynomialFeatures
from sklearn.preprocessing import OneHotEncoder
onehot_encoder=OneHotEncoder(sparse=False)
X = np.array([[1, 3, -2], [-4, 0, -1], [3, 1, 8], [2, 1, 6], [8, 4, 6]])
y = np.array([1, 1, 2, 3, 3])
X1 = np.array([[1, 1, 3, -2], [1, -4, 0, -1], [1, 3, 1, 8], [1, 2, 1, 6], [1, 8, 4, 6]])
reshaped = y.reshape(len(y), 1)
Y_onehot = onehot_encoder.fit_transform(reshaped)
#print(det(X.T@X))
#print(det(X@X.T))
print(Y_onehot)
left_inv = inv(X1.T @ X1) @ (X1.T)
#right_inv = X.T @ inv(X @ X.T)
#print(right_inv)
w = left_inv @ Y_onehot
print(w)

Xnew = np.array([1, -2, 4])
Xnew1 = np.array([1, 1, -2, 4])
Ynew = Xnew1@w
print(Ynew)
```

Results:

```
[[1. 0. 0.]
 [1. 0. 0.]
 [0. 1. 0.]
 [0. 0. 1.]
 [0. 0. 1.]]
[[-0.01005025 -6.03015075  7.04020101]
 [-0.24874372 -1.74623116  1.99497487]
 [ 0.44723618  3.34170854 -3.78894472]
 [ 0.03015075  1.09045226 -1.12060302]]
[ -1.03266332 -10.09798995  12.13065327]
```

7. **MCQ7.1** Suppose your machine learning algorithm achieves 95% accuracy on your training set, but only 10% accuracy on your test set. Which of the following

modification might potentially improve your algorithm's test accuracy? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
(1) Decrease Regularization
(2) Increase Regularization
(3) Decrease Model Complexity
(4) Increase Model Complexity
(5) Decrease Number of Features
(6) Increase Number of Features

Ans: Overfitting regime, so (2) Increase Regularization, (3) Decrease Model Complexity, (5) Decrease number of features

8. **MCQ8.1** Suppose we are minimizing a cost function C(w) with respect to w using a gradient descent algorithm. We observe the following: At iteration 1, C(w) is 11. At iteration 2, C(w) becomes 10.9. At iteration 3, C(w) becomes 10.7. At iteration 4, C(w) becomes 10.6. At iteration 5, C(w) becomes 10.55. Which of the following is true? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
   (1) Decreasing the learning rate will speed up the optimization further
   (2) Increasing the learning rate will speed up the optimization further
   (3) Adding a regularization to the cost function will speed up the optimization further
   (4) Insufficient information in the question to tell which option is true

Ans: (4) Even though the cost function is decreasing slowly, it's entirely possible that we were lucky and initialized near the local minimum, so hard to say without further experimentation.

9. **MCQ8.2** Consider the cost function C(w) = DataLoss(w) + λRegularization(w). Assume that the global minimum of C(w) when λ = 2 is 12. Now we change λ to be equal to 20 and again minimize C(w). Assuming we attain the global minimum, the new optimal cost function value C
   (1) will be higher than before
   (2) will be lower than before
   (3) will stay the same
   (4) Insufficient information in the question to tell which option is true.

Ans: (4) Insufficient information in the question to tell which option is true.

10. **MCQ8.3** Consider the cost function C(w) = DataLoss(w) + λRegularization(w). When λ = 3, the minimum of C(w) is achieved at w = w* and C(w*) = 25. Now we change λ to be equal to 30 and minimize C(w). The minimum now is achieved at w = w**. Assume that we always achieve the global minimum. Which of the following is true? Note that there might be more than one true option. If so, you should select all the correct options.
    (1) In general, Regularization(w*) > Regularization (w**)
    (2) In general, Regularization (w*) < Regularization (w**)
    (3) In general, DataLoss(w*) > DataLoss(w**)
    (4) In general, DataLoss(w*) < DataLoss(w**)
    (5) Insufficient information in the question to tell which option is true.

Ans: By increasing regularization hyperparameter, we are putting more weight on the regularization term, so (1) Regularization (w*) > Regularization (w**) and less weight on the data-loss term, so (4) DataLoss(w*) < DataLoss(w**). Note that one special case is that the Regularization and DataLoss have the same global minimum, which is why I added "in general" qualifier in the statements.

11. **MCQ9.1** Suppose in a 2-class classification problem, our decision tree algorithm achieves 52% accuracy on our training set, and also 52% accuracy on our test set. Which of the following modification might potentially improve our algorithm's test accuracy? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
    (1) Increase maximum depth of tree
    (2) Decrease maximum depth of tree
    (3) Increase minimum number of samples for splitting a leaf node
    (4) Decrease minimum number of samples for splitting a leaf node
    (5) Try random forest instead of decision tree

Ans: Underfitting regime, so (1) Increase maximum tree depth and (4) Decrease minimum number of samples for splitting a leaf node. Although we are in the underfitting regime, but since we are increasing model complexity, we can use try random forest to decrease variance, so (5) is true as well.

12. **MCQ10.1** Which of the following statement(s) about the training-validation-test is/are true? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

    (a) We should always set the number of training samples to be equal to the number of test samples.
    (b) We should always set the number of test samples to be equal to the number of validation samples.
    (c) If no validation data is available, we should test the performances of different models on the test set, and choose the model that performs the best to be our model for deployment.
    (d) None of the others

Ans: (d)

13. **MCQ10.2** Which of the following statement(s) is/are true about the confusion matrix in a classification task? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

    (a) A reasonable binary classifier should always lead to the following result: True Negative + True Positive > False Negative + False Positive.
    (b) A reasonable binary classifier should always leads to the following result: False Negative > False Positive
    (c) A reasonable binary classifier should always leads to the following result: True Positive > True Negative
    (d) None of the others.

Ans: (a). (a) has to be true since, if True Negative + True Positive < False Negative + False Positive, it means the binary classifier is doing worse than random.

14. **MCQ10.3** Which of the following statement(s) is/are true about ROC curve? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

    (a) ROC curve is a widely used evaluation metric.
    (b) Area Under the Curve (AUC) is lower-bounded by 0 and upper-bounded by 1.
    (c) Given a fixed dataset, a higher Gini coefficient of ROC curve usually indicates better performance.
    (d) None of the others

Ans: (a)(b)(c)

15. **MCQ11.1** Which of the following statement(s) is/are true about K-means? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

    (a) Before convergence, K-means is guaranteed to not increase the objective values.
    (b) K-means always produces the same clusters and centroids for different initializations.
    (c) Given a fixed K value, a set of samples, and the objective of K-means, i.e., $J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|\mathbf{x}_i - \mathbf{c}_k\|^2$, if we traverse all possible values of w_{ik} for all i and k, we will find the global optimal solution, which is guaranteed to be no worse than the result of K-means.
    (d) None of the others

Ans: (a)(c)

16. **MCQ11.2** Which of the following statement(s) is/are true about unsupervised learning? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

    (a) In unsupervised learning, we may use the labels of the validation data for reference.
    (b) K-means always yields better results than fuzzy C-means.
    (c) K-means never finds the global optimal solutions.
    (d) None of the others.

Ans: (d)

17. **MCQ12.1** Which of the following statement(s) is/are true about deep neural networks? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

(a) If we build a multilayer perceptron network, we do not need non-linear activation functions.
(b) Backpropagation involves a backward stage but not a forward stage.
(c) Parameters of the neural networks are chosen manually in a smart way, in order to produce the outputs as similar as the labels.
(d) None of the others.

Ans: (d)

Fill in the Blank (In total 9 Questions with 26 blanks, each 2.5 questions, in total 2.5*26 = 65 marks)

1. **FIB6.1** You are given a collection of 6 training data points of one feature (x) and their class label (y) as follows:
$x \in \{4, 7, 10\}$ are labelled as $y = -1$, $x \in \{2, 3, 9\}$ are labelled as $y = +1$.
Answer each of the following subquestions regarding the training outcome or prediction. Note that the training outcome is in terms of the error count, i.e., the total number of data points being incorrectly classified. **[4 Blanks]**

(i) What is the best training error count that you can achieve on these data from a linear classifier, i.e., $f(x)$ = sign(ax + b), on the original input features?

(ii) You need to predict the class label of a new data point given by $x_t = 6$ when a linear classifier, i.e., $f(x)$ = sign(ax + b) is used for prediction. What is the class label (-1 or 1) for x_t?

(iii) What is the best training error count that you can achieve from a 4th order polynomial model, e.g., $f(x)$ = sign($\sum_{k=0}^{4} c_k x^k$)?

(iv) You need to predict the class label of a new data point given by $x_t = 6$ when a 4th order polynomial model, e.g., $f(x)$ = sign($\sum_{k=0}^{4} c_k x^k$) is used for prediction. What is the class label (-1 or 1) for x_t?

Ans:
Original data plot: there are 6 data points altogether, 3 belonging to class 2 ($+1$) and 2 belonging to class 1 (-1). They are not separable.

(i) The linear classifier gives a training error count of 2.

The system can be written as $\mathbf{Xw} = \mathbf{y}$ where $\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & 7 \\ 1 & 10 \\ 1 & 2 \\ 1 & 3 \\ 1 & 9 \end{bmatrix}$, $\mathbf{w} \in \mathcal{R}^{2\times 1}$ and $\mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ +1 \\ +1 \\ +1 \end{bmatrix}$.

Solving the **over-determined** linear system gives $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and the following result:

w: [0.74468085 -0.12765957]
The resulting y for all training samples are
[0.23404255 -0.14893617 -0.53191489 0.4893617 0.36170213 -0.40425532]

After sign → [[1], [-1], [-1], [1], [1], [-1]]
Original y = [-1 -1 -1 1 1 1]
Error count = 2

ii) Predict new y for x= 6
x= [1, 6]
New y = wx = [-0.0212766]
New y class = -1

(iii) The 4th polynomial classifier gives a training error count of 0.

The system can be written as $\mathbf{Pw} = \mathbf{y}$ where $\mathbf{P} = \begin{bmatrix} 1 & 4 & 4^2 & \dots & \mathbf{4}^4 \\ 1 & 7 & 7^2 & \dots & 7^4 \\ 1 & 10 & 10^2 & \dots & 10^4 \\ 1 & 2 & 2^2 & \dots & 2^4 \\ 1 & 3 & 3^2 & \dots & 3^4 \\ 1 & 9 & 9^2 & \dots & 9^4 \end{bmatrix} \in \mathcal{R}^{6\times5}$, $\mathbf{w} \in \mathcal{R}^{5\times1}$

and $\mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ +1 \\ +1 \\ +1 \end{bmatrix}$.

Solving the **over-determined** polynomial system gives $\hat{\mathbf{w}} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y}$ and the following result:

w:
[-11.21971831 13.52136821 -4.88054326   0.65537223 -0.02923541]
Predicted **y**
[-0.76338028 -1.11830986 -1.04225352 1.07605634 0.74647887 1.10140845]
Predicted **y** class
[[-1], [-1], [-1], [1], [1], [1]]
Error count = 0

iv) Predict new y for x= 6
x should be transformed to P first
[[1.000e+00 6.000e+00 3.600e+01 2.160e+02 1.296e+03]]
New y = wP = [-2.11975855]
Newy class = -1

```
#code
import numpy as np
from numpy.linalg import inv
from sklearn.preprocessing import PolynomialFeatures
# linear regression
m_list = [[1, 4], [1, 7], [1, 10], [1, 2], [1, 3], [1, 9]]
X = np.array(m_list)
leftinv_X = np.linalg.inv(X.T @ X) @ X.T
y = np.array([-1, -1, -1, 1, 1, 1])
w = leftinv_X.dot(y)
print(w)
predict_y = X@w
print(predict_y)
```

```
y_class_predict = [[1 if x >=0 else -1 ] for x in predict_y ]
print(y_class_predict)

Xnew = [[1, 6]]
Ynew = Xnew@w
print(Ynew)

## polynomial regression
print('poly2')
## Quadratic 2nd order poly
order = 2
origX= np.array([[4],[7], [10], [2], [3], [9]])
poly = PolynomialFeatures(order)
P = poly.fit_transform(origX)
w_poly =  inv(P.T @ P) @ P.T @ y
print(w_poly)
predict_yp=P@w_poly
print(predict_yp)
yp_class_predict = [[1 if x >=0 else -1 ] for x in predict_yp ]
print(yp_class_predict)

## Order 4 poly: overdetermined
print('poly4')
order = 4
poly = PolynomialFeatures(order)
P = poly.fit_transform(origX)
w_poly = inv(P.T @ P) @P.T @ y
print(w_poly)
predict_yt=P@w_poly
print(predict_yt)
yt_class_predict = [[1 if x >=0 else -1 ] for x in predict_yt ]
print(yt_class_predict)

Xnew = [[1, 1]]
Pnew =poly.fit_transform([[6]])
print(Pnew)
Ynew_poly = Pnew@w_poly
print(Ynew_poly)
```

2. **FIB7.1** Suppose our training set comprises 5 data samples shown in the table below. Each data point consists of 3 features (x1, x2, x3) and a regression target y.

x1	3.3459	1.0893	3.2103	1.744	1.6762
x2	2.7435	2.9113	1.4706	1.2895	2.1366
x3	-1.7253	-0.7804	-0.9944	0.5307	-1.0502
y	2.9972	1.1399	2.228	0.3387	2.5042

On the other hand, our test set comprises 4 data samples shown in the table below.

x1	0.9478	1.4931	2.5809	2.0607
x2	1.1619	1.7705	2.9569	1.1695
x3	0.7779	-0.6149	-1.2087	0.4972
y	0.4168	1.4707	2.1037	0.9177

Our goal is to train a quadratic regression model to predict y. Suppose we utilize Pearson's correlation to perform feature selection.

For the purpose of feature selection, the correlation between feature x1 and target y is equal to BLANK1 (your answer should be up to 3 decimal places)

For the purpose of feature selection, the correlation between feature x2 and target y is equal to BLANK2 (your answer should be up to 3 decimal places)

For the purpose of feature selection, the correlation between feature x3 and target y is equal to BLANK3 (your answer should be up to 3 decimal places)

Based on the correlations, the best feature is feature BLANK4 (Your answer should be an integer 1, 2 or 3 corresponding to feature 1, feature 2 or feature 3).

[4 Blanks]

Ans:

BLANK 1: Corr between x1 and y: 0.65099
BLANK 2: Corr between x2 and y: 0.37222
BLANK 3: Corr between x3 and y: -0.93081
BLANK 4: 3 (because we should look at highest absolute correlation)

See matlab code below.

```
training = [3.3459   1.0893   3.2103   1.7440   1.6762;
   2.7435   2.9113   1.4706   1.2895   2.1366;
   -1.7253  -0.7804  -0.9944   0.5307  -1.0502;
   2.9972   1.1399   2.2280   0.3387   2.5042]

disp(['Corr between x1 and y: ' num2str(corr(training(1, :)', training(4, :)'))]);
disp(['Corr between x2 and y: ' num2str(corr(training(2, :)', training(4, :)'))]);
disp(['Corr between x3 and y: ' num2str(corr(training(3, :)', training(4, :)'))]);
```

3. **FIB8.1** We would like to minimize the cost function $C(w) = \sin^2(w)$ using gradient descent. Suppose we initialize our algorithm with w = 3. The gradient of the cost function at w = 3 is BLANK1 (your answer should contain up to 3 decimal places). Suppose the step size is 0.1. After the first round of gradient descent, the updated value of w is BLANK2 (your answer should be up to 3 decimal places). **Note that w is in radians.**

[2 Blanks]

Ans:

Gradient is given by $2\sin(w)\cos(w)$

At w = 3, gradient = 2*sin(3)*cos(3) = **-0.279415498**

After first round of gradient descent, updated value of w = 3 **-** (-0.279415498)*0.1 **= 3.027942**

4. **FIB8.2** We would like to minimize the cost function $C(x, y) = x^2 + xy^2$ using gradient descent. Suppose we initialize our algorithm with x = 3, y = 2. Suppose the step size is 0.2. After the first round of gradient descent, the updated value of x is BLANK1 (your answer should be up to 2 decimal places) and the updated value of y is BLANK2 (your answer should be up to 2 decimal places).

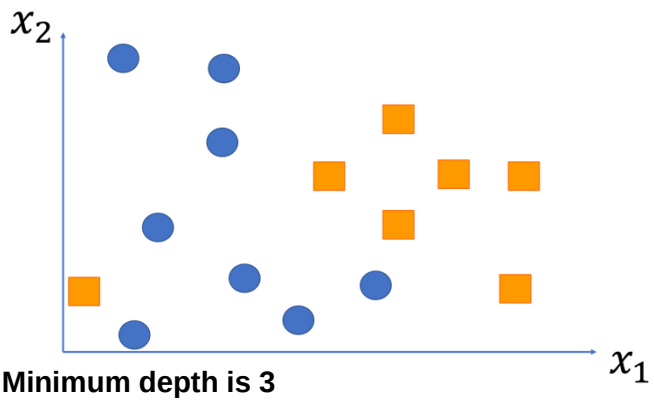
[2 Blanks]

Gradient wrt x is given by $2x + y^2$.
At x = 3, y = 2, gradient = 2(3) + 2^2 = 10.
After first round of gradient descent x = 3 - 0.2*10 = **1.00**

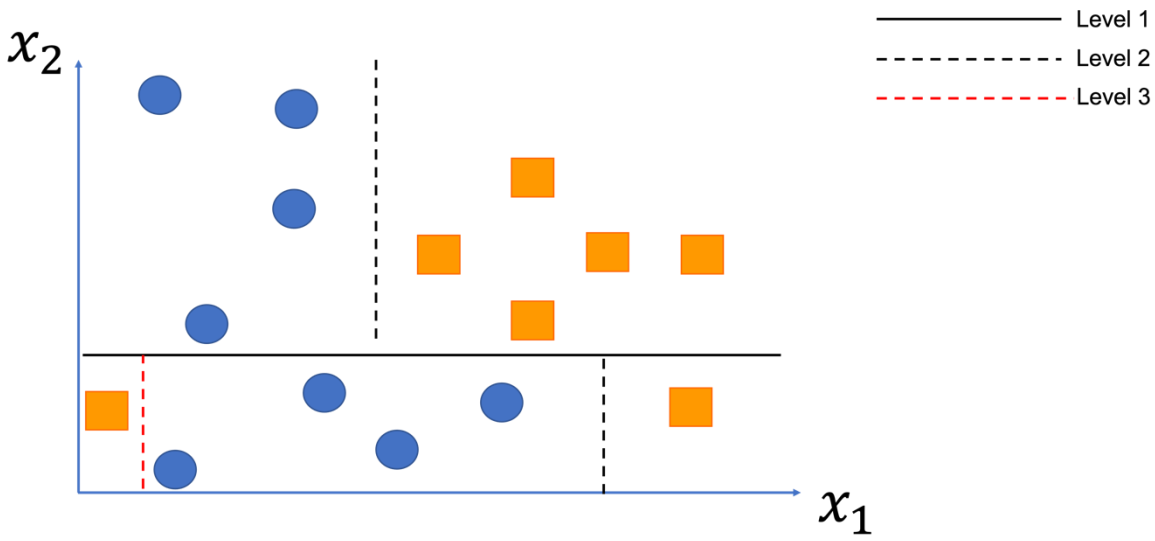
Gradient wrt y is given by 2xy.
At x = 3, y = 2, gradient = 2(3)(2) = 12
After first round of gradient descent, y = 2 - 0.2*12 = **-0.40**

5. **FIB9.1** To classify the data points in the plot below with 100% accuracy, a decision tree needs to have a minimum depth of BLANK (your answer should be a positive integer)

[1 Blank]



Ans: Minimum depth is 3



6. **FIB9.2** Consider the following dataset comprising 10 datapoints: {x, y} = {0.2, 2.1}, {0.7, 1.5}, {1.8, 5.8}, {2.2, 6.1}, {3.7, 9.1}, {4.1, 9.5}, {4.5, 9.8}, {5.1, 12.7}, {6.3, 13.8}, {7.4, 15.9}. Our goal is to use a regression tree to predict y from x. Suppose at depth 1, we consider a decision threshold of 3.

The MSE at the root is BLANK1 (your answer should be up to 3 decimal places).

The overall MSE at depth 1 is BLANK2 (your answer should be up to 3 decimal places).

[2 Blanks]

Ans:
MSE at root = 20.6381
Overall MSE at depth 1 = 5.5648

See matlab code below.

```
function RegressionTreeExam

rng(2);

x = [0.2 0.7 1.8 2.2 3.7 4.1 4.5 5.1 6.3 7.4];
y = x*2 + round(rand(1,length(x))*40)/10;
threshold = 3;

disp([x; y])
disp(['threshold = ' num2str(threshold)]);
disp(' ');
```

```
% root
root_mse = mean((y - mean(y)).^2);
disp(['root mse = ' num2str(root_mse)]);

% left x < threshold
yL = y(x < threshold);
numL = length(yL);
mse_L  = mean((yL - mean(yL)).^2);
disp(['(x < threshold) mse = ' num2str(mse_L)]);

% right x > threshold
yR = y(x > threshold);
numR = length(yR);
mse_R  = mean((yR - mean(yR)).^2);
disp(['(x > threshold) mse = ' num2str(mse_R)]);

% Overall MSE
overall_MSE = (mse_L * numL + mse_R * numR)/length(y);
disp(['overall mse = ' num2str(overall_MSE)]);
```

7. **FIB 10.1**
   In a binary classification task, we use a dataset that contains in total 1000 samples, where we take out 200 samples as the test set. We then conduct a standard 4-fold cross validation. In each fold, we set the number of validation samples to be the same as that of test samples. As such, in each fold, the binary classifier is trained on *_BLANK1_* samples.

   Now we have three classifier candidates, where their average training, validation, and number of parameters are given as follows. We should adopt classifier #*_BLANK2_* for the test set.

Classifiers	Average Training Accuracy	Average Validation Accuracy	Number of parameters
Classifier #1	90.5%	92.1%	5000
Classifier #2	91.5%	92.1%	3000
Classifier #3	93.5%	90.5%	3000

[2 Blanks]

Ans:
BLANK1: 600
BLANK2: 2

For BLANK1, Number of training samples = 1000*(1-2*20%) = 600

For BLANK1, Classifier #2 has the highest validation accuracy and fewer number of parameters, and a higher training accuracy as compared to Classifier #1.

8. **FIB 10.2**
Jack developed a classifier to detect spam emails. The classifier conducts binary classification, and identifies an email to be spam (positive class) or not spam (negative class). The detected spam emails go into the 'Spam' folder and non-spam emails go into the 'Inbox' folder.

In an unseen dataset of 100 emails, the classifier achieves the following performances

	Spam (Predicted)	Not Spam (Predicted)
Spam (Actual)	40	*x*
Not Spam (Actual)	15	35

The number *x is _BLANK1_*.

We can also derive that, among the 100 emails, *_BLANK2_* emails have the actually label of being Spam, and in total *_BLANK3_* emails are correctly classified.

Assume we have the following cost matrix

	Spam (Predicted)	Not Spam (Predicted)
Spam (Actual)	$C_{s,s}$ * 40	$C_{s,n}$ * *x*
Not Spam (Actual)	$C_{n,s}$ * 15	$C_{n,n}$ * 35

We assume that, if an email is correctly classified, we impose no penalty, i.e., $C_{s,s} = C_{n,n} = 0$. On the other hand, we know that customers will be very unhappy if the spam folder contains non-spam emails. But customers find it acceptable that some spam emails go into the inbox folder. In this case, when comparing two classifiers, we should ensure that $C_{s,n}$ *_BLANK4_* $C_{n,s}$ (choose '=', '>' or '<' and fill in the blank).

With the same dataset, David also developed a classifier that achieves the following performance

	Spam (Predicted)	Not Spam (Predicted)
Spam (Actual)	50	*y*
Not Spam (Actual)	10	25

With the same set of $\{C_{n,n}, C_{s,s}, C_{n,s}, C_{s,n}\}$ as above, David's classifier is considered to be *_BLANK5_* (1:Better or 2:Worse; choose '1' or '2' and fill in the blank) than the one of Jack's.

[5 Blanks]

Ans:
BLANK1: 10
BLANK2: 50

BLANK3: 75
BLANK4: <
BLANK5: 1

9. **FIB 11.1**

We have a collect of 10 books. We measure their thickness in millimetres and summarize them as follows

Book ID	01	02	03	04	05	06	07	08	09	10
Thickness (mm)	50	60	66	68	71	72	75	82	90	99

We'd like to group the books into two groups according to their thickness.

1) Assume we pick book 03 as the initial centroid for Group A, book 07 for Group B, and assign the books to the two groups using Euclidean distance. Before updating the new centroid, we will have *_BLANK1_* books in Group A (please enter an integer here)

2) With the above group assignments, we re-estimate the new centroids for the two groups. The new centroid of Group A is *_BLANK2_* , and the new centroid of Group B is *_BLANK3_* . (please specify your answer with two decimal places)

3) With updated centroid, we run group assignment again. In this second round, the number of books in the Group A (the group with lower centroid) is *_BLANK4_*.

[4 Blanks]

Ans

BLANK1: 4
BLANK2: 61
BLANK3: 81.5
BLANK4: 5

END OF PAPER