[2]First name :	Last name :	Grade :

University of M'sila
Faculty of Mathematics and Computer Science
Departement of Computer Science

جامعة المسيلة
كلية الرياضيات والإعلام الآلي
قسم الإعلام الآلي

Exam

Course : Machine Learning Models **Academic Year : 2023 - 2024**

Exercise 1 (12 pts)

1. What do we call a straight line that attempts to predict the relationship between two points?

A. Regression line
B. Scatter Plot
C. Line segment
D. Slope

2. In simple linear regression, if you change the input value by 1 then output value will be changed by:

A. 1
B. The slope parameter
C. The intercept parameter
D. None

3. What would be the best regression model for more than one independent variable?

A. Simple Linear Regression
B. Multiple Linear Regression
C. Logistic Regression
D. All of the Above

4. Suppose you have observed that you data has an exponential growth tendency. Then what regression model you should use ?

A. Simple linear regression
B. Multiple linear regression
C. Polynomial regression
D. Logistic regression

5. If you get a poor accuracy using a simple linear regression model. What will be the cause behind it ?

A. The data was not linear
B. The data has outliers
C. Both A or B depending on the context
D. None

6. Suppose you got a training accuracy of 90% and a test accuracy of 50%. What happened with your model ?

A. The model was over fitted with the training data
B. The model was under fitted with the training data
C. The model is absolutely fine
D. None

7. Which one is a different algorithm?

A. Logistic Regression
B. Support Vector Regression
C. Polynomial Regression
D. None

8. Suppose you have to predict the salary of employees from their experience. This is a :

A. Classification task
B. Regression task
C. Clustering task
D. None

9. Which of the following is an example of an ensemble method?

A. Support vector machines
B. K Nearest Neighbors
C. K-means
D. Random forests
E. None

10. What is underfitting in supervised learning?

A. A model that is too complex and has learned the noise in the training data
B. A model that performs well on new, unseen data
C. A model that is too simple and does not capture the underlying patterns in the data
D. A model that performs well on both training and testing data

11. What is overfitting in supervised learning?

A. A model that performs poorly on training data
B. A model that performs poorly on new, unseen data
C. A model that performs well on both training and testing data
D. A model that is too complex and has learned the noise in the training data
E. None

12. A classifier :

A. Inputs a vector of continuous values and outputs a single discrete value
B. Inputs a vector of discrete values and outputs a single discrete value
C. Both A and B
D. None

13. What does k stand for in the KNN algorithm?

A. Number of output classes

B. Number of output clusters

C. Number of neighbors

D. Number of input features

E. None

14. What is bagging in ensemble learning?

A. Training multiple base models on the same data and features

B. Training multiple base models on different subsets of the data and features

C. Training multiple base models on different subsets of the data and using a weighted average to combine their predictions

D. Training a meta-model to combine the predictions of multiple base models

15. What is the main difference between a hard-margin SVM and a soft-margin SVM?

A. A hard-margin SVM is computationally efficient, while a soft-margin SVM is computationally expensive

B. A hard-margin SVM allows no classification errors, while a soft-margin SVM allows some classification errors

C. A hard-margin SVM can handle missing data, while a soft-margin SVM cannot

16. k-NN algorithm does more computation on test time rather than train time.

A. TRUE

B. FALSE

17. Figures 1 and 2 illustrate decision boundaries for two nearest-neighbour classifiers.

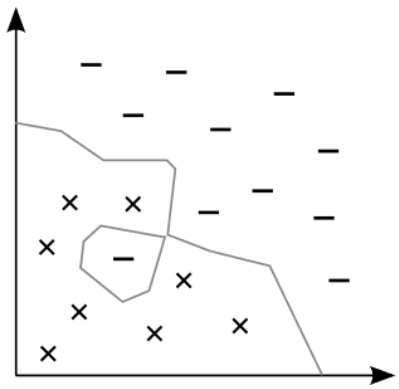


Figure 1

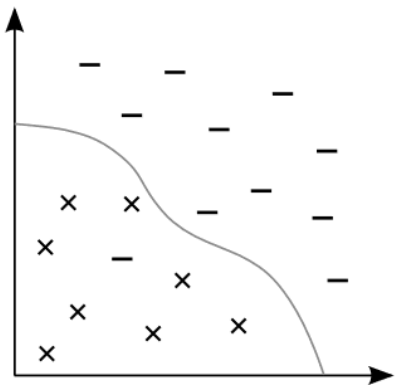


Figure 2

Determine which one of the boundaries belongs to the 1-nearest neighbour classifier and which one belongs to the 3-nearest neighbour classifier.

A. Figure 1 represents the boundaries to the 1-NN and Figure 2 represents the boundaries to the 3-NN.

B. Figure 2 represents the boundaries to the 1-NN and Figure 1 represents the boundaries to the 3-NN.

C. Can’t say.

18. When you find noise in data which of the following option would you consider in k-NN?

A. I will increase the value of k
B. I will decrease the value of k
C. Noise cannot be dependent on value of k
D. Apply another algorithm
E. None of these

19. How do you handle imbalanced data in the KNN algorithm?

A. Weighted voting, where the vote of each neighbor is weighted by its inverse distance to the query point.
B. Oversample the minority class.
C. Undersample the majority class.
D. All of the above.

20. Suppose you have a dataset with n = 10 features and m = 5000 examples. After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets.

Which of the following might be promising steps to take? Select all that apply.

A. Increase the bias
B. Use an SVM with a Gaussian Kernel.
C. Create / add new polynomial features.
D. Use an SVM with a linear kernel, without introducing new features.
E. Reduce the number of example in the training set.
F. None of these

21. Which of the following statements are true? Select all that apply.

A. Since K-Means is an unsupervised learning algorithm, it cannot overfit the data, and thus it is always better to have a larger number of clusters.
B. If we are worried about K-Means getting stuck in bad local optima, one way to solve this problem is by using multiple random initializations.
C. For some datasets, the “right” or “correct” value of K can be ambiguous and hard even for a human expert looking carefully at the data to decide.
D. The standard way of initializing K-Means is by setting all centroid of clusters to vectors of zeros.

22. Which of the following statements is true about Ensemble Learning?

A. It always performs worse than individual models.
B. It is a combination of unrelated models.
C. It can combine models of the same type only.
D. It can improve model accuracy and generalization.

23. You run K-Means 100 times with different initialization. How should pick from the 100 resulting solutions?

A. Pick randomly –that was the point of random initialization.
B. Pick the one with the lowest cost.
C. Average all 100 solutions together.
D. Pick the last one because K-Means always improves over time.

24. What is the purpose of using a loss function in supervised learning?

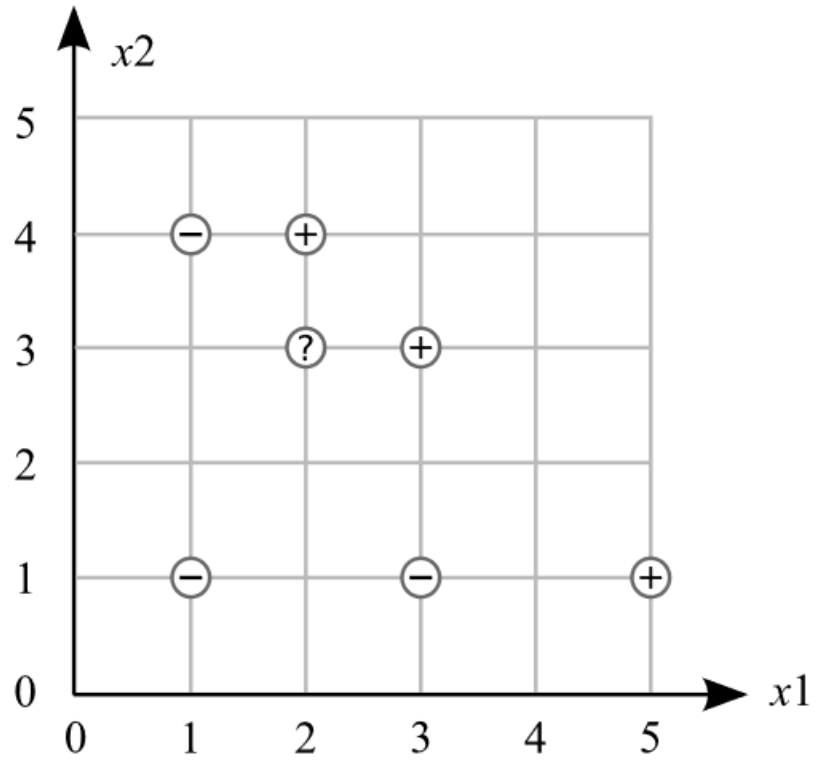
A. To identify input features

B. To measure the error between predicted and actual outputs

C. To optimize hyperparameters

D. To find the best model architecture

Exercise 2 (02 pts)



Data points are: Negative: (1, 1) (3, 1) (1, 4) Positive: (2, 4) (3, 3) (5, 1). Data points are classified as either +1 or -1. An unknown point is located at (2, 3)

Questions:

1) Draw the 1-NN decision boundaries on the graph above.

2) How would 1-NN classify the unknown point (2, 3).

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Exercise 3 (06 pts)

Train an SVM model that can predict the class for the labelled dataset “data.csv” containing 3 features and the last column contains the label values (class name is “class”). Complete the following python code:

 # Import libraries

```
import pandas as pd
import numpy as np
```

……

……

……

 # Load the CSV file (C:\data.csv): Store the result into a dataframe variable called data.

……

……

 # Separate the class labels (Y) from the features (X)

……

……

 # Separation train set (80%) from test set (20%)

……

……

 # Create an SVM classifier model

```
SVM_Model = svm.SVC(gamma=0.5, C=0.1)
```

 # Train the model

……

……

 # Make prediction for the test set

……

……