Tuesday, November 16, 2021 6:07 PM

GIT INTRO QUESTIONS
GIT ADD,commit,push

LABS 1-7 REVIEW THOSE QUESTIONS

l is Data Science?

Data Science is an interdisciplinary field that involves the use of statistical and computati
insights from data.

What is a PIVOT Table?
How do we use pivot tables to helps us do
data visualization
Answer business questions
feature engineer?
Aggregate Functions

What are the steps involved in a Data Science project?
Be aware what the code looks like for each of the machine Learning steps

- Define the problem
  - How do we identify a birt using a computer
- Gather data
  - Bird pictures
- Preprocess the data
  - Label our images extract key components into features
-
- Create a model
- Train Model
- Evaluate the model

onal methods to extract

- Evaluate the model
- Deploy the model
- Monitor the performance of the model
- Repeat until desired performance is reached

BE ABLE TO IDENTIFY WHAT STEP IS DOING WHAT IN SCIKITLEARN

What is Machine Learning?

Answer: Machine Learning is a subset of Artificial Intelligence that involves the use of alg
models to enable machines to learn from data and make predictions or decisions without
programmed.
No if else etc…

What are the different types of Machine Learning algorithms?

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

What is the difference between supervised and unsupervised learning?

In supervised learning, the algorithm learns from labeled data, whereas in unsupervised
learns from unlabeled data.

What is classification in Machine Learning?

Answer: Classification is a type of supervised learning algorithm that involves predicting a
outcome variable based on one or more input variables.\

What is regression in Machine Learning? Answer: Regression is a type of supervis
that involves predicting a continuous or numerical outcome variable based on on
variables.

What is clustering in Machine Learning? Answer: Clustering is a type of unsupervised le
involves grouping similar data points together based on their characteristics.

What is feature engineering? Answer: Feature engineering is the process of selecting a
input variables or features used in a machine learning model to improve its performanc

What is overfitting in Machine Learning?

Answer: Overfitting occurs when a machine learning model is too complex and fits the tra
performs poorly on new or unseen data.

orithms and statistical
t being explicitly

learning, the algorithm

a categorical or discrete

ed learning algorithm
e or more input

earning algorithm that

nd transforming the
ce.

aining data too well, but

What is cross-validation?

Answer: Cross-validation is a technique used to assess the performance of a machine lea the data into training and testing sets multiple times and averaging the results.

What is hyperparameter tuning?

Answer: Hyperparameter tuning is the process of selecting the best set of hyperparamet algorithm to optimize its performance.

What is an ROC curve?

Answer: An ROC (Receiver Operating Characteristic) curve is a graphical representation of the true positive rate and the false positive rate of a machine learning model at different thresho

What is an AUC score?

1. **Answer: AUC (Area Under the Curve) is a metric used to evaluate the performance of model based on the area under the ROC curve.**
2. **What is a confusion matrix?**
3. **Answer: A confusion matrix is a table that shows the number of true positives, false positives false negatives of a machine learning model.**

What is the difference between a training set and a testing set? Answer: A training set machine learning model, whereas a testing set is used to evaluate the performance of

What is a decision tree?

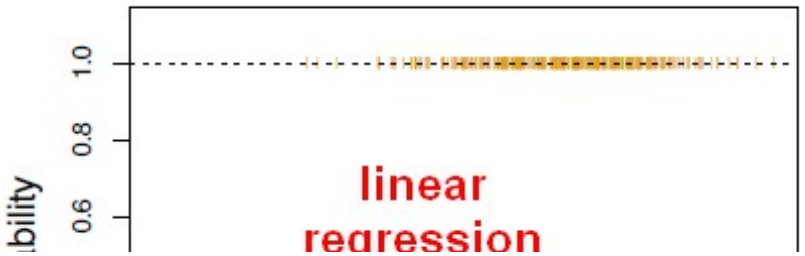
Answer: A decision tree is a tree-like model used in machine learning to make decisions rules or conditions.

What is random forest?

Answer: Random forest is an ensemble learning method that uses multiple decision trees to reduce the variance of a machine learning model.

What is logistic regression?

is a type of regression analysis used to predict the probability of a binary or categorical outco one or more input variables.



rning model by splitting

ers for a machine learning

e trade-off between the olds.

a machine learning

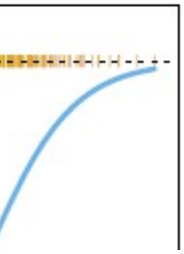
s, true negatives, and

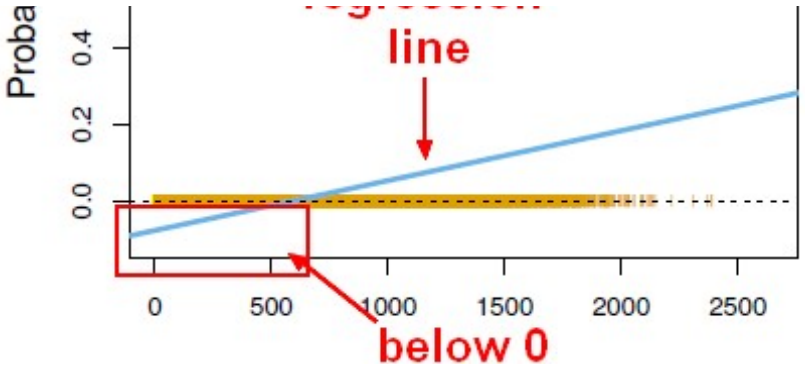
is used to train a the model.

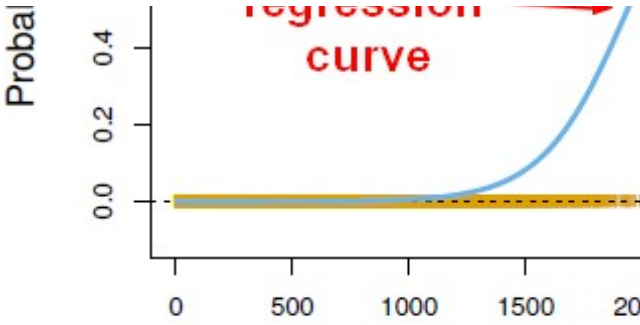
s based on a series of

improve the accuracy and

ome variable based on







What is K-means clustering?

K-means clustering is a type of unsupervised learning algorithm used to partition a set clusters based on their characteristics.

What is support vector machine (SVM)?

Support vector machine (SVM) is a supervised learning algorithm used to classify data by find maximally separates the classes.

What is a neural network?

A neural network is a set of interconnected nodes or neurons that can learn from data decisions.

What is NumPy?

What is the difference between Pandas Series and Pandas DataFrame?

Answer: Pandas Series is a one-dimensional labeled array that can hold any data type, whereas two-dimensional labeled data structure that can hold multiple Pandas Series

SERIES

DATAFRAME2D MATRIX

SERIES	SERIES	SERIES

of data points into K

ding the hyperplane that

to make predictions or

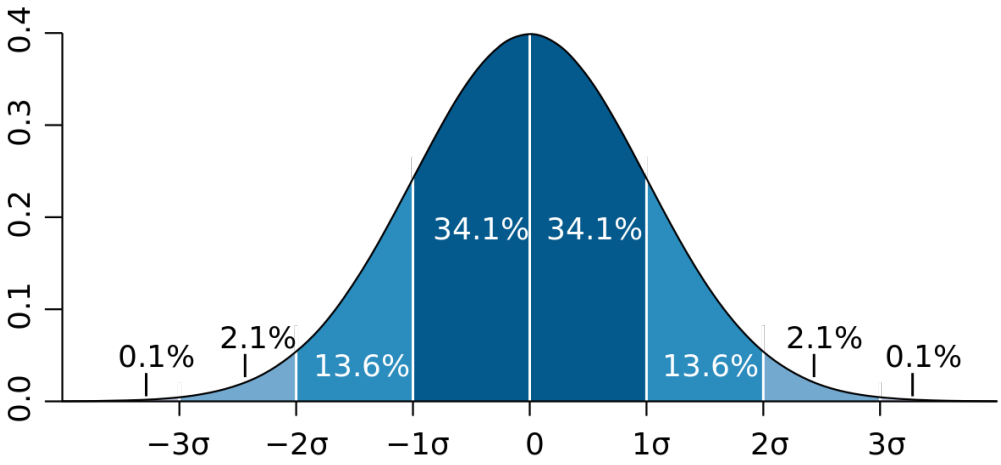
Pandas DataFrame is a

What is the difference between a Python list and a NumPy array?
Answer: NumPy arrays are **more efficient an**d provide more **functionality f**or numerical compu
which are more general-purpose data structures.

What is broadcasting

Broadcasting is a feature in NumPy that allows arrays of different shapes to be used in ar
automatically aligning their dimensions.

What is the difference between variance and standard deviation? Answer: Variance is a
spread out a set of data is from its mean, calculated by taking the average of the squa
mean, while standard deviation is the square root of the variance and provides a meas
data in the same units as the original data.



Standard Deviation

76	84	69	92	58
89	73	97	85	77

$$\sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

$$\bar{X} = \frac{\text{Sum}}{n}$$

uting than Python lists,

ithmetic operations by

a measure of how
ed differences from the
ure of the spread of the

- Supervised Learning: Supervised learning is a
  algorithm learns from labeled data, which m
  during training. The goal of supervised learn
  to output variables. In other words, the algo
  from the input variables based on the labele
- Common Algorithms in Supervised Learning:
  supervised learning include:
  - Linear Regression
  - Logistic Regression
  - Decision Trees
  - Random Forest
  - Support Vector Machines (SVM)
  - K-Nearest Neighbors (KNN)
  - Naive Bayes
  - Neural Networks
- Classification vs. Regression:
  - Classification involves predicting a cate
    predicting whether an email is spam or
  - Regression involves predicting a contin
    predicting house prices based on featu
    bedrooms.
- Feature Selection: Feature selection is the pr
  features or variables from the original set of

a type of machine learning where the
eans it is provided with input-output pairs
ing is to learn a mapping from input variables
rithm learns to predict the output variable
d training data.
: Some common algorithms used in

egorical output variable. For example,
r not spam.
uous output variable. For example,
res like size, location, and number of

rocess of selecting a subset of relevant
features to use in the model. The objective

What is the purpose of data normalization?
; Dat
large

11. In Scikit-learn, what is the purpose of the predict method? a) To train the model us provided training data b) To make predictions on new data c) To evaluate the perfo the model d) To visualize the model's decision boundaries
Answer: b) To make predictions on new data

12. Which Scikit-learn function is commonly used to split a dataset into training and tes train_test_split b) split_data c) test_train_split d) dataset_split
Answer: a) train_test_split
13. What is the recommended ratio for splitting a dataset into training and testing sets 80:20 c) 60:40 d) It depends on the size of the dataset and the problem domain
Answer: d) It depends on the size of the dataset and the problem domain
14. What is the purpose of setting a random seed when splitting a dataset into training sets? a) To ensure reproducibility of the results b) To increase the model's accuracy decrease the model's variance d) To speed up the training process
Answer: a) To ensure reproducibility of the results
15. What is the difference between training error and testing error? a) Training error m performance of the model on the training data, while testing error measures its per on unseen data b) Training error measures the performance of the model on unsee while testing error measures its performance on the training data c) There is no diff both terms refer to the same concept d) Training error measures the time taken to model, while testing error measures the time taken to test it
Answer: a) Training error measures the performance of the model on the training d testing error measures its performance on unseen data
16. What is overfitting in the context of machine learning models? a) When the model well on the training data but poorly on unseen data b) When the model performs p both the training and testing data c) When the model has too few parameters to ca underlying patterns in the data d) When the model converges too quickly during tra
Answer: a) When the model performs well on the training data but poorly on unsee
17. How can overfitting be mitigated in machine learning models? a) By increasing the of the model b) By decreasing the size of the training dataset c) By regularizing the using techniques like cross-validation d) By training the model for fewer epochs
Answer: c) By regularizing the model or using techniques like cross-validation
18. What is underfitting in the context of machine learning models? a) When the mode

ard variables with
ing the
rmance of

sting sets? a)

? a) 70:30 b)

and testing
c) To

neasures the
rformance
en data,
ference;
train the

lata, while

performs
oorly on
pture the
aining
en data
complexity
model or

el performs

of feature selection is to improve the model
decreasing computational cost, and enhanci

- Feature Standardization: Feature standardiz
process of scaling or normalizing the range o
transforming the features so that they have
one. Feature standardization is needed beca
do not dominate those with smaller scales d
converge faster and improves the stability o
- Overfitting: Overfitting occurs when a mode
noise and random fluctuations in the data ra
to poor generalization performance on unse
  - Using more training data
  - Cross-validation to evaluate model per
  - Feature selection to reduce model com
  - Regularization techniques such as L1 a
  - Early stopping during model training
- Underfitting: Underfitting occurs when a mo
structure of the data. This results in poor pe
unseen data. Underfitting can be prevented
  - Increasing model complexity (e.g., add
  - Adding more features to the model
  - Reducing regularization constraints
  - Using a more complex model algorithm

's performance by reducing overfitting,
ng model interpretability.
ation, also known as feature scaling, is the
of features in the dataset. It involves
a mean of zero and a standard deviation of
use it ensures that features with larger scales
uring the training process. It helps algorithms
f the models.
l learns the training data too well, capturing
ather than the underlying patterns. This leads
en data. Overfitting can be prevented by:

formance
nplexity
nd L2 regularization

odel is too simple to capture the underlying
rformance both on the training data and
by:
ing more layers to a neural network)

n