| **Study Abroad: Data Science and AI with Python. Retake Exam** | Marks obtained ↓ |
|---|---|
| Date: 18.01.2024, Total questions: **27** Total points: **75** | |
| Name: Time: 1.5 hrs | |

| Question: | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Points: | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Score: | | | | | | | |
| Question: | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Points: | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Score: | | | | | | | |
| Question: | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| Points: | 1 | 1 | 1 | 1 | 1 | 1 | 5 |
| Score: | | | | | | | |
| Question: | 22 | 23 | 24 | 25 | 26 | 27 | Total |
| Points: | 5 | 5 | 10 | 10 | 10 | 10 | 75 |
| Score: | | | | | | | |

## 1 Multiple choice questions

**Instructions:**

1. **All questions count 1 point**
2. **Wrong answers substract 0.25 points**

[1] 1. In Python, which library is commonly used for data manipulation and analysis?

A. TensorFlow

B. Scikit-Learn

**C. Numpy**

D. PyTorch

[1] 2. What is the primary advantage of using Jupyter notebooks for data analysis in Python?

A. Better performance

**B. Code modularity**

C. Real-time collaboration

D. Strong typing

[1] 3. In statistical learning, what does the term ”Supervised Learning” mean?

**A. Learning with labeled data**

B. Learning with a teacher or mentor

C. Learning without guidance

D. Learning using neural networks

[1] 4. What is the main goal of clustering in unsupervised learning?

A. Predicting outcomes

B. Finding hidden patterns

**C. Minimizing error**

D. Reducing dimensionality

1 5. Which of the following clustering algorithms is based on centroid initialization and iterative assignment and update steps?

A. DBSCAN

B. Agglomerative Hierarchical Clustering

C. Mean Shift

**D. K-Means**

1 6. In hierarchical clustering, what is the linkage criterion used to measure the distance between clusters when merging them?

A. Euclidean distance

B. Manhattan distance

C. Silhouette coefficient

**D. Ward's method**

1 7. What is a limitation of the K-Means clustering algorithm?

A. Sensitivity to cluster shapes

**B. Sensitivity to outliers**

C. Difficulty handling noisy data

D. Lack of scalability

1 8. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is known for:

A. Partitioning data into a predetermined number of clusters

B. Agglomerative merging of clusters

**C. Clustering data based on density connectivity**

D. Assigning data points to the nearest centroid

1 9. What does PCA stand for in the context of dimensionality reduction?

**A. Principal Component Analysis**

B. Primary Component Analysis

C. Predictive Cluster Algorithm

D. Probability and Causality Analysis

1 10. What does the term ”Unsupervised learning” mean?

A. Learning without any feedback

B. Learning with a mentor

C. Learning using neural networks

**D. Learning without labeled data**

1 11. What is the primary difference between linear and nonlinear models in regression?

A. Linearity of the relationship

B. Number of features

**C. Degree of complexity**

D. Use of optimization techniques

1 12. What role do kernel models play in machine learning?

A. Dimensionality reduction

B. Clustering of data points

C. Regularization of models

**D. Nonlinear transformation of input features**

1 13. What is KNN (K-Nearest Neighbors) primarily used for in classification?

A. Dimensionality reduction

**B. Assigning labels based on nearest neighbors**

C. Clustering

D. Predicting probabilities

1 14. How does SVM (Support Vector Machines) handle nonlinear decision boundaries?

A. By increasing the model complexity

B. By reducing the number of support vectors

C. By using gradient boosting

**D. By transforming features using kernels**

1 15. What is the primary advantage of using decision trees in machine learning?

A. Robustness to outliers

B. High computational efficiency

**C. Simplicity and interpretability**

D. Ability to handle high-dimensional data

1 16. In ensemble methods, how does Random Forest improve model performance?

A. By reducing bias

**B. By combining multiple weak learners**

C. By increasing variance

D. By removing outliers

1 17. How are categorical variables typically encoded in decision trees?

A. Label Encoding

**B. One-Hot Encoding**

C. Integer Encoding

D. Binary Encoding

1 18. What is the primary difference between a neural network and a traditional machine learning model?

A. Use of linear models

B. Lack of optimization techniques

**C. Depth and complexity**

D. Independence from data size

1 19. In deep learning, what is the purpose of Convolutional Neural Networks (CNN)?

A. Sequence prediction

B. Clustering

C. Dimensionality reduction

**D. Image classification and recognition**

1 20. What does the term ”Training NN” refer to in the context of deep learning?

**A. Fine-tuning model parameters**

B. Normalizing input data

C. Evaluating model performance

D. Reducing model complexity

## 2 Regular questions

**Instructions:**

1. **Give clear answers**
2. **Use only the space provided below the question**

5 21. Consider a dataset with the following points: A(2, 3), B(5, 4), C(3, 6), D(8, 7), E(9, 5). Perform the first iteration of the K-Means clustering algorithm with initial centroids at A(2, 3) and D(8, 7).

**Solution:**

1. Initial centroids: A(2, 3) and D(8, 7)
2. Assignment: A, B, C to cluster 1; D, E to cluster 2
3. Update centroids: (3.33, 4.33) and (8.5, 6)

5 22. Compare and contrast K-Means clustering and hierarchical clustering algorithms. Highlight their differences in terms of approach, scalability, and sensitivity to cluster shapes.

**Solution:**

- K-Means: Centroid-based, iterative, scalable, sensitive to cluster shapes.
- Hierarchical Clustering: Tree-based, not as scalable, captures hierarchical relationships, less sensitive to cluster shapes.

5 23. Suppose you have a dataset with outliers, and you want to use a clustering algorithm that is less sensitive to outliers. Which clustering algorithm would you choose, and why?

**Solution:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) would be a good choice because it can identify clusters of varying shapes and is less sensitive to outliers. It defines clusters as dense regions separated by sparser regions, making it robust against outliers.

10 24. Explain the concept of Principal Component Analysis (PCA) in the context of dimensionality reduction. How does PCA work, and what is the primary goal of this technique in machine learning?

**Solution:** Principal Component Analysis (PCA) is a dimensionality reduction technique used in machine learning and statistics. Its primary goal is to transform high-dimensional data into a lower-dimensional representation while retaining as much of the original variability as possible. PCA achieves this by identifying and selecting the principal components, which are linear combinations of the original features.

The steps involved in PCA are as follows:

- Normalization: Standardize the data to have zero mean and unit variance.
- Covariance Matrix: Compute the covariance matrix of the standardized data.
- Eigendecomposition: Find the eigenvectors and eigenvalues of the covariance matrix.

- Principal Components: Sort the eigenvectors by their corresponding eigenvalues in descending order to obtain the principal components.
- Projection: Project the original data onto the selected principal components to obtain a reduced-dimensional representation.

PCA is widely used for reducing the dimensionality of datasets, eliminating redundant information, and improving computational efficiency in machine learning models.

10 25. Explain the concept of a Random Forest in machine learning. What are the key components of a Random Forest, and how does it contribute to the overall performance of the algorithm? Additionally, how does a Random Forest handle overfitting compared to a single decision tree?

**Solution:** A Random Forest is an ensemble learning method in machine learning that operates by constructing a multitude of decision trees during training and outputs the mode of the classes for classification tasks or the average prediction for regression tasks.

Key components of a Random Forest include:

- Decision Trees: The base learner in a Random Forest is a decision tree. However, instead of using a single decision tree, a forest is constructed by training multiple trees.
- Random Subspace Sampling: During the training process, each decision tree in the forest is trained on a random subset of the features, known as the random subspace. This introduces diversity among the trees.
- Bootstrap Aggregating (Bagging): Each tree in the forest is trained on a bootstrapped sample of the original dataset. Bagging helps in reducing variance and improving the generalization of the model.
- Voting or Averaging: For classification tasks, the final prediction is determined by a majority vote from all the trees, while for regression tasks, it is the average prediction across all trees.

The Random Forest addresses overfitting compared to a single decision tree through the following mechanisms:

- Diversity: The use of random subspace sampling and bagging introduces diversity among the trees, making the overall model more robust and less prone to overfitting.
- Ensemble Averaging: The combination of predictions from multiple trees helps to smooth out individual tree idiosyncrasies and reduces the risk of capturing noise in the data.
- Out-of-Bag (OOB) Error: Random Forests use an out-of-bag error estimate during training, which is calculated on data not used in the bootstrap sample. This provides an unbiased estimate of the model's performance and helps in monitoring for overfitting.

In summary, a Random Forest is a powerful ensemble learning method that leverages the strength of multiple decision trees, introduces randomness to improve generalization, and effectively addresses overfitting concerns.

10 26. Explain the key concepts and components of Convolutional Neural Networks (CNNs) in the context of image processing. What is the role of convolutional layers, pooling layers, and fully connected layers in a typical CNN architecture? Also, how do CNNs capture hierarchical features in images?

**Solution:** Convolutional Neural Networks (CNNs) are a class of deep neural networks designed for tasks related to image processing and computer vision. The key components of CNNs include convolutional layers, pooling layers, and fully connected layers.

- Convolutional Layers: Convolutional layers apply convolution operations to the input data. These operations involve sliding small filters (kernels) across the input image to extract local features. The filters learn to detect patterns like edges, textures, and simple shapes. Multiple filters are used to capture various features.
- Pooling Layers: Pooling layers downsample the spatial dimensions of the feature maps produced by convolutional layers. Common pooling operations include max pooling and average pooling. Pooling helps reduce the dimensionality of the data, making it computationally efficient. It also introduces a form of translation invariance.
- Fully Connected Layers: Fully connected layers process the high-level features learned by the previous layers and make final predictions. These layers connect every neuron to every neuron in the previous and subsequent layers, forming a traditional neural network structure.

CNNs capture hierarchical features in images through the arrangement of convolutional layers:

Early layers detect simple features like edges and textures. Intermediate layers combine simple features to detect more complex patterns and shapes. Deep layers represent high-level features and object compositions. CNNs leverage parameter sharing and local connectivity in convolutional layers, enabling them to learn hierarchical representations efficiently. The shared weights in convolutional layers allow the network to recognize the same pattern regardless of its location in the input, contributing to the model's translation invariance.

In summary, CNNs use convolutional layers to extract local features, pooling layers to downsample and enhance translation invariance, and fully connected layers for high-level feature processing. The hierarchical structure of CNNs enables them to learn and represent complex visual features in images.

10 27. Can you briefly explain the different steps in this code?

```
import tensorflow as tf
mnist = tf.keras.datasets.mnist
(train_images, train_labels), (test_images, test_labels) = mnist.load_data()
train_images, test_images = train_images / 255.0, test_images / 255.0
model = tf.keras.Sequential([
        tf.keras.layers.Flatten(input_shape=(28, 28)),
        tf.keras.layers.Dense(128, activation='relu'),
        tf.keras.layers.Dense(10, activation='softmax')
        ])
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])
model.fit(train_images, train_labels, epochs=5)
test_loss, test_acc = model.evaluate(test_images, test_labels)
print("Test accuracy:", test_acc)
```

**Solution:**

Listing 1: TensorFlow Pseudocode for Simple Neural Network

```
import tensorflow as tf

# Load MNIST dataset
```

```
mnist = tf.keras.datasets.mnist
(train_images, train_labels), (test_images, test_labels) = mnist.load_data()

# Normalize pixel values
train_images, test_images = train_images / 255.0, test_images / 255.0

# Build the neural network model
model = tf.keras.Sequential([
        tf.keras.layers.Flatten(input_shape=(28, 28)),
        tf.keras.layers.Dense(128, activation='relu'),
        tf.keras.layers.Dense(10, activation='softmax')
        ])

# Compile the model
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

# Train the model
model.fit(train_images, train_labels, epochs=5)

# Evaluate the model
test_loss, test_acc = model.evaluate(test_images, test_labels)
print("Test accuracy:", test_acc)
```