Introduction to Machine Learning Mid-Term Exam

Multiple Choice: Fundamental Concepts

1. What is the main goal of supervised learning?

1. To find hidden patterns or structures in unlabeled data.
2. To classify data into predefined categories or predict numerical values.
3. To group similar data points together based on their features.
4. To optimize a function by iteratively adjusting model parameters.

2. Which of the following is an example of unsupervised learning?

1. Image classification.
2. Sentiment analysis.
3. Clustering customer data.
4. Predicting stock prices.

3. What is the purpose of feature scaling in machine learning?

1. To convert categorical variables into numerical ones.
2. To remove outliers from the dataset.
3. To normalize the range of feature values.
4. To increase the interpretability of the model.

Multiple Choice: Algorithms

4. Which algorithm is suitable for handling large-scale datasets with high-dimensional features?

1. Decision Trees.
2. Logistic Regression.
3. Support Vector Machines.
4. Naive Bayes.

5. Which algorithm can be used for both classification and regression tasks?

1. K-Nearest Neighbors.
2. Random Forest.
3. AdaBoost.
4. Gradient Boosting.

6. Which algorithm is known for its ability to handle imbalanced datasets and has applications in fraud detection?

1. K-Means Clustering.
2. Gaussian Mixture Models.
3. Principal Component Analysis.
4. Support Vector Machines.

Short Answer: Evaluation Metrics

7. Define accuracy, precision, recall, and F1 score in the context of binary classification.

Accuracy: the ratio of the correctly classified data to the total amount of classifications made by the model.

Precision: quantifies the number of positive class predictions that actually belong to the positive class.

Recall: quantifies the number of positive class predictions made out of all positive examples in the dataset.

F-Measure: provides a single score that balances both the concerns of precision and recall in one number.

8. What are the advantages and disadvantages of using accuracy as an evaluation metric?

Advantages: looks at fractions of correctly assigned positive and negative classes. When your problem is balanced, using accuracy is usually a good start. An additional benefit is that it is really easy to explain to a non-technical stakeholders in your project,

Disadvantages: A common complaint about accuracy is that it fails when the classes are imbalanced. Ithas issues when the classes are naturally balanced, too. In many applications, there are different costs associated with the different mistakes. Accuracy takes away from your ability to play the odds. The typical threshold for a (binary) model that outputs probability values (logistic regression, neural nets, and others) is 0.5. Accuracy makes a 0.49 and 0.51 appear to be different categories while 0.51 and 0.99 are the same.

9. Explain the concept of overfitting and underfitting in machine learning models.

Overfitting means our training has focused on the particular training set so much that it has missed the point entirely. In this way, the model is not able to adapt to new data as it’s too focused on the training set.

Underfitting means the model has not captured the underlying logic of the data. It doesn’t know what to do with the task we’ve given it and, therefore, provides an answer that is far from correct.

True or False: Evaluation Metrics

10. The area under the ROC curve (AUC-ROC) is a commonly used evaluation metric for imbalanced datasets. T

11. Precision is the ratio of true positive predictions to the sum of true positive and false positive predictions. T

12. The F1 score is the harmonic mean of precision and recall, and it provides a balanced measure between the two. T

Short Answer: Cross-Validation

13. Explain the concept of cross-validation and its importance in evaluating machine learning models.

Cross-validation is a popular technique in machine learning used to evaluate the performance of a predictive model. It is used to prevent overfitting in a predictive model, especially when the available data is insufficient. Understanding cross-validation enables practitioners to make informed decisions and improve the effectiveness of machine learning algorithms.

14. What is k-fold cross-validation, and how does it work?

K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. It involves splitting the data into k equal-sized subsets, or folds, and using one fold as the test set and the remaining k-1 folds as the training set. This process is repeated k times, each time using a different fold as the test set. K-fold cross-validation helps to estimate the model's ability when given new data.

15. Name one potential drawback of using cross-validation.

Most cross-validation methods involve performing many tests. Each of these tests takes time to perform. Some cross-validation methods, specifically the exhaustive type, can take a lot of time to complete. If you plan on performing cross-validation, consider scheduling extra time to perform your tests to help finish them before your deadline.