Midterm Exam
Machine Learning

Tập hợp câu hỏi và đáp án chi tiết

Tổng số câu hỏi:	10 câu
Thời gian:	Không giới hạn
Dạng bài:	Trắc nghiệm & Tự luận
Ngày tạo:	25/08/2025

Câu hỏi 1

Which of the following statements accurately describes a key distinction between Machine Learning (ML) and traditional programming?

A. Traditional programming focuses on developing computer programs that can learn from data without being explicitly programmed, while ML requires explicit rules.

B. Machine Learning takes data and desired outcomes to generate a program (model), whereas traditional programming uses data and explicit rules to produce output. ✓

C. Machine Learning is primarily concerned with manual feature engineering, while traditional programming automates feature extraction.

D. Traditional programming is a subset of Artificial Intelligence that deals with data analysis, while Machine Learning is a subset of Statistics.

Đáp án đúng: B

Giải thích:
As stated in 'Introduction to Machine Learning' (Course Material 2), Machine Learning (ML) takes data and desired outcomes to generate a program (model), allowing computers to learn without being explicitly programmed. In contrast, traditional programming relies on explicit rules applied to data to produce output.

Câu hỏi 2

A data scientist is analyzing customer behavior for an online store. They observe that 20% of customers who visit the website make a purchase. Among those customers who make a purchase, 75% also sign up for the newsletter. What is the probability that a randomly selected customer who visits the website both makes a purchase AND signs up for the newsletter?

A. 0.95

B. 0.15 ✓

C. 0.55

D. 0.20

Đáp án đúng: B. 0.15

Giải thích:
This problem requires the application of the Product Rule of probability. Let P be the event that a customer makes a purchase, and N be the event that a customer signs up

for the newsletter. We are given P(P) = 0.20 (20% of customers make a purchase) and P(N|P) = 0.75 (75% sign up for the newsletter GIVEN they made a purchase). We need to find the joint probability P(P and N). According to the Product Rule, P(P and N) = P(N|P) * P(P). Therefore, P(P and N) = 0.75 * 0.20 = 0.15. This avoids the common mistake of confusing joint, marginal, and conditional probabilities, as the 'among those' phrase clearly indicates a conditional probability.

Câu hỏi 3

In the context of training a linear regression model using Batch Gradient Descent, which statement accurately describes the algorithm's primary objective and the role of the learning rate (α)?

A. To minimize the Mean Squared Error (MSE) cost function J(w) by iteratively updating the model parameters (w); a larger α can lead to faster convergence but risks overshooting the minimum or diverging. ✓

B. To maximize the accuracy of predictions on unseen data by adjusting the model's complexity; a smaller α ensures the algorithm finds the global minimum quickly.

C. To directly calculate the optimal parameters (w) in a single step using the normal equation; α determines the number of features included in the model.

D. To increase the variance of the model to better fit complex datasets; a larger α helps prevent overfitting by adding a penalty to large parameter values.

Đáp án đúng: A. To minimize the Mean Squared Error (MSE) cost function J(w) by iteratively updating the model parameters (w); a larger α can lead to faster convergence but risks overshooting the minimum or diverging.

Giải thích:
Gradient Descent is an iterative optimization algorithm whose primary objective is to find the parameters (w) that minimize the cost function J(w), which for linear regression is typically the Mean Squared Error (MSE). The learning rate (α) controls the size of each step taken during the parameter updates. A larger α can indeed lead to faster convergence by taking bigger steps, but it also carries the risk of overshooting the minimum point of the cost function or causing the algorithm to diverge entirely. Conversely, a very small α would lead to very slow convergence.

Câu hỏi 4

Which of the following statements correctly differentiates between Maximum A Posteriori (MAP) and Maximum Likelihood Estimation (MLE) and accurately describes a core characteristic of the Naïve Bayes classifier?

A. MAP estimation is primarily used when there is no prior knowledge about the hypothesis, whereas MLE incorporates prior beliefs, and Naïve Bayes is a discriminative model.

B. MLE is a special case of MAP where all hypotheses are assumed to have equal prior probabilities, and the Naïve Bayes classifier addresses the curse of dimensionality by assuming conditional independence of features. ✓

C. Naïve Bayes uses MAP estimation to overcome the curse of dimensionality by directly modeling the joint probability P(x,y) without relying on feature independence.

D. Both MAP and MLE are iterative optimization methods for discriminative models, while Naïve Bayes is a generative model that requires a closed-form solution for its parameters.

Đáp án đúng: B

Giải thích:
Option B is correct. According to Course Material 1, MLE is a special case of MAP where all hypotheses are assumed to have equal prior probabilities (P(h) is constant), meaning MLE maximizes only the likelihood P(D|h). Course Material 2 states that the Naïve Bayes classifier addresses the 'Curse of Dimensionality' by assuming conditional independence of features, simplifying $P(x|C_k)$ to $\Pi(j=1 \text{ to } d)\ P(x_j|C_k)$. Option A incorrectly describes the roles of prior knowledge in MAP and MLE and misclassifies Naïve Bayes. Option C incorrectly states that Naïve Bayes uses MAP to overcome the curse of dimensionality without relying on feature independence; in fact, it relies heavily on this independence. Option D incorrectly categorizes MAP and MLE as iterative optimization methods for discriminative models and misrepresents the parameter estimation for Naïve Bayes.

Câu hỏi 5

What is the primary goal of using impurity measures such as Gini Index and Entropy in decision tree algorithms?

A. To determine the overall accuracy of the decision tree on unseen data.

B. To calculate the computational complexity of building the tree.

C. To identify attribute splits that lead to child nodes with more homogeneous class distributions. ✓

D. To ensure that all branches in the tree have an equal number of records.

Đáp án đúng: C

Giải thích:
Impurity measures like Gini Index and Entropy are fundamental to decision tree induction. Their primary purpose is to evaluate potential splits by quantifying the homogeneity of class distributions within nodes. The goal is to select splits that result in 'purer' child nodes, meaning nodes where the records belong predominantly to a single class, as stated in Course Material 2.

Câu hỏi 6

Explain the Bias-Variance Tradeoff in machine learning. In your response, define bias and variance, describe how each contributes to a model's generalization error, and clarify their respective roles in causing underfitting and overfitting.

Đáp án mẫu:

Bias represents the error from an overly simplified model, leading to underfitting where the model consistently misses true values and exhibits high empirical and true error; it contributes to approximation error. Variance represents the error from a model's excessive sensitivity to the training data, leading to overfitting where it learns noise and performs inconsistently on new data, contributing to estimation error. The Bias-Variance Tradeoff dictates that increasing model complexity typically reduces bias but increases variance, and vice-versa. Minimizing generalization error (true error on unseen data) requires finding an optimal balance between these two, as both high bias (underfitting) and high variance (overfitting) result in poor performance.

Câu hỏi 7

Explain the fundamental difference in the probabilistic modeling approach between a generative classifier like Naïve Bayes and a discriminative classifier like Logistic Regression. Subsequently, discuss how this core difference impacts their respective advantages and disadvantages, specifically concerning their performance characteristics and data requirements.

Đáp án mẫu:

Generative models like Naïve Bayes learn the joint probability distribution P(x, y) or P(x|y) and P(y) to infer P(y|x) using Bayes' Theorem, while discriminative models like Logistic Regression directly model the conditional probability P(y|x) or the decision boundary. This means Naïve Bayes can perform well with less data due to stronger assumptions (e.g., feature independence) and has closed-form solutions for parameters, but its performance can degrade if these assumptions are violated. Conversely, Logistic Regression generally performs better with sufficient training data and is more robust to noisy data as it directly optimizes for classification, though it lacks a closed-form solution and requires iterative, potentially slower, optimization methods.

Câu hỏi 8

Describe the fundamental operational principles of K-Means and DBSCAN clustering algorithms. Explain how their distinct approaches lead to different strengths and weaknesses when dealing with datasets containing non-globular cluster shapes or significant noise, providing an example of a data characteristic where one algorithm clearly outperforms the other.

Đáp án mẫu:

K-Means is a partitioning algorithm that iteratively assigns data points to the nearest centroid, aiming to minimize the Sum of Squared Error (SSE) and requiring the number of clusters (K) beforehand. In contrast, DBSCAN is a density-based algorithm that defines clusters as high-density regions, identifying core, border, and noise points based on parameters Eps and MinPts. K-Means struggles with non-globular cluster shapes and is sensitive to noise and outliers, as they can distort centroids. DBSCAN, however, excels at discovering arbitrary-shaped clusters and explicitly identifies noise points, making it superior for datasets with irregular cluster geometries and significant outliers.

Câu hỏi 9

A data science team at 'PropTech Innovations' is developing a linear regression model to predict housing prices based on various features like living area (sqft), number of bedrooms, and location score. They have collected a dataset of 500 housing records, each with 3 input features (x1, x2, x3) and a continuous target variable (y). The team decides to use a multivariable linear regression model, h(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3, and the Mean Squared Error (MSE) as

their cost function. **Scenario 1: Optimization Algorithm Choice** The team is debating between two general approaches for finding the optimal parameters (w): a direct analytical solution or an iterative optimization algorithm. Given the dataset size (m=500) and the nature of the problem, they lean towards an iterative approach. **Scenario 2: Hyperparameter Tuning** After implementing Batch Gradient Descent, the team observes the following behaviors during training: * **Attempt A**: The cost function J(w) initially decreases but then starts to increase rapidly and eventually becomes 'NaN' (Not a Number). * **Attempt B**: The cost function J(w) decreases very slowly, taking an extremely long time to converge, even after thousands of iterations. **Scenario 3: Model Performance Evaluation** After training a model that appears to converge, the team evaluates its performance on both the training dataset and a separate, unseen test dataset. They observe two distinct outcomes for different model configurations: * **Configuration X**: The model achieves a very low MSE on the training dataset (e.g., near zero), but a significantly higher MSE on the test dataset. * **Configuration Y**: The model achieves a high MSE on both the training dataset and the test dataset. **Your Task:** 1. **For Scenario 1:** Justify why Gradient Descent is a suitable iterative optimization algorithm for this problem, explaining its core mechanism based on the provided course materials. (Approx. 2-3 sentences) 2. **For Scenario 2:** * Diagnose the likely cause for the observed behavior in **Attempt A** and **Attempt B** regarding the cost function's convergence. * For each attempt, propose a specific adjustment to the learning rate (alpha) that would likely resolve the issue, explaining *why* that adjustment is appropriate based on Gradient Descent principles. 3. **For Scenario 3:** * Identify the specific problem (e.g., overfitting, underfitting) indicated by the performance in **Configuration X** and **Configuration Y**. * For each configuration, propose a mitigation strategy based on the concepts introduced in the course materials. Explain how your proposed strategy addresses the identified problem.

Đáp án mẫu:

1. **For Scenario 1:** Gradient Descent is suitable because it iteratively updates the model parameters (w) by moving in the direction of the steepest decrease of the cost function J(w). It uses the partial derivatives of the MSE cost function with respect to each parameter to determine this direction, allowing it to find the optimal parameters that minimize the prediction error for the given dataset. This iterative approach is particularly effective for larger datasets where direct analytical solutions might be computationally expensive or infeasible. 2. **For Scenario 2:** * **Attempt A (Cost becomes NaN):** This behavior indicates that the learning rate (alpha) is too large. A large alpha causes the parameter updates to overshoot the minimum of the cost function, leading to oscillations of increasing magnitude or divergence, eventually causing the cost to explode and become 'NaN'. * **Attempt B (Slow convergence):** This behavior indicates that the learning rate (alpha) is too small. A small alpha results in very tiny steps during each parameter update, causing the algorithm to take an

excessively long time to reach the minimum of the cost function. * **Adjustment for Attempt A:** Decrease the learning rate (alpha). A smaller alpha will ensure smaller steps, preventing overshooting and allowing the algorithm to converge smoothly towards the minimum. * **Adjustment for Attempt B:** Increase the learning rate (alpha). A larger alpha will allow the algorithm to take bigger steps, accelerating convergence towards the minimum, provided it doesn't become too large and cause divergence. 3. **For Scenario 3:** * **Configuration X (Low training MSE, high test MSE):** This indicates **overfitting**. The model has learned the training data too well, including its noise and specific patterns, but fails to generalize to unseen data. * **Configuration Y (High MSE on both training and test):** This indicates **underfitting**. The model is too simple or has not learned enough from the training data, resulting in poor performance on both the training and unseen test sets. * **Mitigation for Configuration X (Overfitting):** Overfitting occurs when the hypothesis fits the training data too well. A mitigation strategy, based on the provided materials, would be to consider if the hypothesis space chosen (e.g., linear model with these specific features) is too complex for the amount of data, or if the model is capturing noise. While not explicitly detailed for linear regression, simplifying the model or ensuring the features are truly representative could be considered. (Note: More advanced techniques like regularization are not covered in the provided materials for Week 3). * **Mitigation for Configuration Y (Underfitting):** Underfitting often results from choosing an inappropriate hypothesis space. To mitigate this, the team should consider if the current linear model ($h(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3$) is too simple to capture the underlying relationship in the data. They might need to explore a more complex hypothesis space, for example, by adding polynomial features (e.g., $x1^2$, $x1*x2$) or interaction terms, to allow the model to fit the data better.