Quiz 1

CSL7620: Machine Learning | AY 2024-25 | Sem I

The respondent's email () was recorded on submission of this form.

✗ **Consider the following scenario for Stochastic Gradient Descent (SGD):** *0/1

You are training a linear regression model using SGD on a single data point. The equation for the output is y=θ(0)+θ(1)x. The learning rate is set to 0.1. Suppose at a certain iteration, the model parameters θ(0) and θ(1) are θ(0)=2.0 and θ(1)=1.5, respectively. The current data point you use to update the model is x=4, and the true label y is 10.

Which of the following are correct statements after one SGD update for the parameters θ(0) and θ(1)?

- ○ New value of θ(0) and θ(1) will be 1.8 and 1.3, respectively.
- ○ New value of θ(0) and θ(1) will be 2.2 and 2.3, respectively.
- ○ New value of θ(0) and θ(1) will be 2.2 and 1.7, respectively.
- ◉ New value of θ(0) and θ(1) will be 1.8 and 0.7, respectively. ✗

Correct answer

- ◉ New value of θ(0) and θ(1) will be 2.2 and 2.3, respectively.

✓ **Using VC-Dimension, we can get an idea about which of the following?** * 1/1

- ◉ A model's capacity of classification ✓
- ○ Probability of some limit of test error of a model
- ○ The size of the dataset required for training
- ○ The number of features in the dataset

✓ **Which of the following statements is/ are usually true about bias and variance?** *1/1

- ◉ High variance usually leads to overfitting ✓
- ○ Increasing the complexity of a model always reduces both bias and variance.
- ○ Bias and variance are independent of each other.
- ○ High bias leads to high variance in a model.

✓ **What is the primary objective of a regression model?** * 1/1

- (●) To find a curve that best fits the data points ✓
- ( ) To predict categorical labels
- ( ) To reduce the number of features in the dataset
- ( ) To cluster similar data points together

✓ **Which of the following techniques are used for data normalization?** * 1/1

- [x] Min-Max Scaling ✓
- [ ] One-Hot Encoding
- [ ] Log Transformation
- [x] Z-score Normalization ✓

✓ **A key characteristic of unsupervised learning is:** * 1/1

- ( ) Predicting future outcomes
- (●) Discovering hidden patterns in data ✓
- ( ) None of the others
- ( ) Learning from rewards

✓ **Which of the following statements is/ are usually true about inductive learning?** *1/1

- [x] Inductive learning models can be evaluated based on their ability to predict outcomes on new, unseen data. ✓
- [ ] Inductive learning can only be applied to supervised learning scenarios.
- [x] Overfitting is a common problem in inductive learning when a model learns the training data too well, including noise. ✓
- [x] Inductive learning assumes that the patterns in the training data will hold true for unseen data. ✓

✗ **Which of the following scenarios usually is indicative of high variance in a model?** *0/1

- [x] Large differences between training and testing accuracy ✓
- [ ] None of the others
- [ ] More inductive bias
- [ ] High VC dimension

Correct answer

- [x] Large differences between training and testing accuracy
- [x] High VC dimension

✓ **Which of the following statements correctly describes how reinforcement learning works?** *1/1

- [ ] Reinforcement learning does not require interaction with the environment to learn optimal actions.
- [ ] The agent learns to make decisions by directly performing the optimal actions without feedback.
- [x] Reinforcement learning involves controlling an agent through a sequence of good decisions. ✓
- [x] The learning process in reinforcement learning is driven by experiences through rewards and penalties from the environment. ✓

✗ **For a given hypothesis space, if the VC dimension is d, which of the following is/are correct?** *0/1

- [ ] None of the others
- [x] The hypothesis space must not consist of only 2D straight lines if d>3 in 2D Euclidean plane ✓
- [x] The hypothesis space can perform correct binary classification for any set of d points ✗
- [x] The hypothesis space cannot shatter any set of d+1 points ✓

Correct answer

- [x] The hypothesis space cannot shatter any set of d+1 points
- [x] The hypothesis space must not consist of only 2D straight lines if d>3 in 2D Euclidean plane

Quiz 2

CSL7620: Machine Learning | AY 2024-25 | Sem I

The respondent's email () was recorded on submission of this form.

✓ **In MLE, we often use the log-likelihood function instead of the likelihood function because** *1/1

- (●) Taking the logarithm simplifies calculations by turning products into sums. ✓
- ( ) The log-likelihood transforms the data to make it normally distributed.
- ( ) The log-likelihood avoids overfitting the model.
- ( ) The log-likelihood is easier to interpret.

✓ **Which of the following is a key advantage of DBSCAN over K-means?** * 1/1

- ( ) DBSCAN does not require any input from the user
- (●) DBSCAN can automatically find the optimal number of clusters ✓
- ( ) DBSCAN is computationally faster than K-means in all cases
- ( ) DBSCAN always works better than K-means with high-dimensional data

✗ **A face recognition system uses a minimum-error classification approach with three classes: Class 1 (known person), Class 2 (unknown person), and Class 3 (mask detection). Given that the misclassification of a known person as an unknown person is a major issue, what among the following will you do to minimize this error while maintaining the overall classification accuracy.** *0/1

- (●) Assign a higher prior probability value to Class 1. ✗
- ( ) Increase the likelihood of Class 3 detections.
- ( ) None of the others
- ( ) Assign a higher loss to misclassifying Class 1 as Class 2.

Correct answer

- (●) Assign a higher loss to misclassifying Class 1 as Class 2.

✓ **Which of the following is/are characteristics of Gaussian Mixture Models (GMMs)?** *1/1

- [x] The mean (μ) and covariance (Σ) are updated iteratively ✓
- [ ] The model minimizes the likelihood of the data
- [x] Each cluster is modeled as a multivariate Gaussian distribution ✓
- [ ] GMMs always use hard clustering to assign points

✗ **In the context of Expectation Maximization, which of the following are correct about the E-step?** *0/1

- [x] It updates the parameters μ and Σ for each Gaussian ✗
- [ ] It assigns hard labels to the data points
- [x] It calculates the probability that each data point belongs to each cluster ✓
- [ ] It uses Bayes' Theorem to calculate responsibilities (λ)

Correct answer

- [x] It calculates the probability that each data point belongs to each cluster
- [x] It uses Bayes' Theorem to calculate responsibilities (λ)

✗ **Which of the following statements is/ are usually true in the context of maximum likelihood estimation (MLE)?** *0/1

- [x] MLE is a technique to minimize the distance between data points. ✗
- [x] Likelihood function measures the probability of observing the data given a set of parameters. ✓
- [ ] In MLE, we never assume that the observations (data points) are independent
- [x] MLE may use gradient descent ✓

Correct answer

- [x] Likelihood function measures the probability of observing the data given a set of parameters.
- [x] MLE may use gradient descent

✓ **Which of the following statements are true about K-means clustering?** * 1/1

- [x] K-means minimizes the sum of squared distances between data points and their corresponding cluster centroid. ✓
- [ ] K-means always converges
- [x] K-means is sensitive to the initial placement of centroids. ✓
- [ ] No input from user is required in K-means

✗ **Which of the following statements is/ are true about Bayes Decision Rule?** *0/1

- [ ] It only works for two-class problems.
- [ ] None of the others
- [ ] It minimizes the probability of classification error.
- [x] It is used to decide the class with the highest posterior probability. ✓

Correct answer

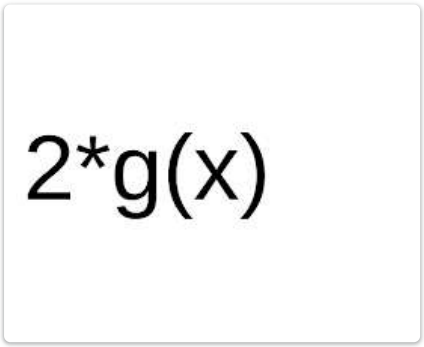
- [x] It minimizes the probability of classification error.
- [x] It is used to decide the class with the highest posterior probability.

✓ **Which of the following is a key assumption of the Naive Bayes classifier?** *1/1

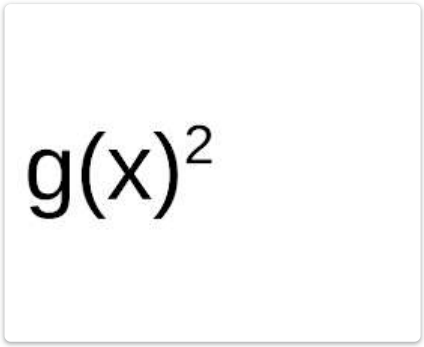
- ( ) The prior probabilities are uniform.
- ( ) None of the others
- ( ) The features are correlated.
- (•) The features are independent given the class label. ✓

✗ **Let g(x) be a discriminant function for a classification task. Then, which of these can also surely be a discriminant function for that task?** *0/1

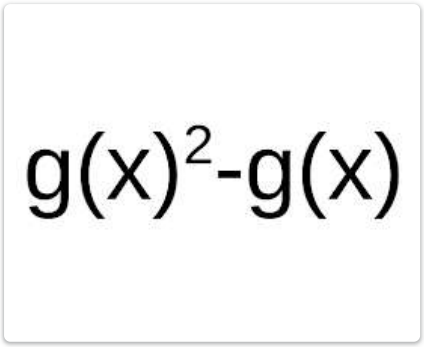
☑ exp(g(x)) ✓



☑ 2g(x) ✓



☑ g(x)^2 ✗



☐ (g(x)^2)-g(x)

Correct answer

☑ exp(g(x))

☑ 2g(x)

Quiz 3

CSL7620: Machine Learning | AY 2024-25 | Sem I

The respondent's email () was recorded on submission of this form.

✓ **In boosting, how does growing a new stump impact previously misclassified samples?** *1/1

- [x] Misclassified samples are given increased weights for growing the next stump. ✓
- [ ] Misclassified samples are excluded from training the next stump.
- [x] The algorithm uses a weighted entropy/ Gini impurity to grow the stumps after the first stump ✓
- [x] Correctly classified samples can be given decreased weights for growing the next stump. ✓

✗ **Find out the correct statements about Support Vector Machine (SVM).** * 0/1

- [ ] Eventually, SVM does not use all the training data points to draw the decision boundary
- [x] SVM tries to maximize the margin between the boundaries of different classes ✓
- [ ] In SVM objective function, we try to minimize misclassification errors
- [ ] SVM can perform only linear classification

Correct answer

- [x] SVM tries to maximize the margin between the boundaries of different classes
- [x] Eventually, SVM does not use all the training data points to draw the decision boundary

✓ **In Linear Discriminant Analysis (LDA), the algorithm focuses on maximizing the separation between classes by:** *1/1

- ( ) Maximizing the variance within each class.
- (•) Maximizing the distance between the mean of each class ✓
- ( ) Minimizing the covariance between features.
- ( ) minimizing the spread within each class.
- ( ) Minimizing the number of principal components needed for classification.

Which of the following statements are true regarding the polynomial kernel in SVM? *0/1

- [x] A higher degree leads to a simpler decision boundary. ✗
- [x] The degree parameter controls the flexibility of the decision boundary ✓
- [ ] It is suitable for data with non-linear boundaries
- [x] It projects data into an infinite-dimensional space. ✗

Correct answer

- [x] It is suitable for data with non-linear boundaries
- [x] The degree parameter controls the flexibility of the decision boundary

Which of the following can be used as a stopping criterion for growing a Decision Tree? *1/1

- [ ] None of the others
- [x] Impurity in child nodes ✓
- [x] Tree depth ✓
- [x] Number of samples in a node ✓

Identify the correct statements * 0/1

- [ ] Eigen values do not play any role in PCA
- [ ] Kernels can be useful in the context of PCA
- [ ] LDA can not be used if we have more than two classes in a dataset
- [x] PCA and ICA are supervised learning techniques ✗

Correct answer

- [x] Kernels can be useful in the context of PCA

Which of the following statements is true regarding the Random Forest algorithm? *1/1

- ( ) Random Forest performs well on small datasets but struggles with large datasets.
- (•) We want the trees in random forest to have low similarities among themselves ✓
- ( ) Random Forest only uses the most important features for training each tree.
- ( ) Each tree in the Random Forest is trained on the entire dataset.

✓ **Which of the following statements about the RBF kernel in SVM is/are correct?** *1/1

- [x] The RBF kernel maps the original data into an infinite-dimensional space ✓
- [ ] The RBF kernel is only useful for linearly separable data.
- [ ] None of the others
- [x] A large gamma in the RBF kernel makes the decision boundary more sensitive to individual data points ✓

✗ **Which of the following statements are true about Decison Tree? *** 0/1

- [x] Decision Trees are used as an unsupervised learning algorithm. ✗
- [ ] Decision trees can not be used for regression.
- [x] A higher value of the Gini Impurity (G) indicates greater heterogeneity within a node. ✓
- [x] Homogeneity of a node generally decreases as we move from parent node to child node. ✗

Correct answer

- [x] A higher value of the Gini Impurity (G) indicates greater heterogeneity within a node.

✓ **When performing Principal Component Analysis (PCA), the principal components are chosen based on:** *1/1

- ( ) The directions that maximize the mean of the projected data on that direction.
- ( ) The directions that minimizes variance of the projected data on that direction.
- ( ) None of the others.
- (•) The directions that maximizes variance of the projected data on that direction. ✓