CS181 Mini-Quiz Questions

Lecture 2: K-Means Clustering

Clustering Variants Which of the following clustering algorithms uses a data example to represent each cluster?

A) Agglomerative Clustering

B) K-Means

C) K-Nearest Neighbors

D) K-Medoids

Answer: [D] K-Medoids is like K-Means, but rather than means, chooses a data example to represent each cluster.

Convergence True or False? Lloyd's algorithm is guaranteed to find the globally optimal solution to K-Means.
Answer: False: Lloyd's algorithm only converges to a local minimum.

Model Selection "Oversegmentation" in clustering refers to

A) Partitioning the data into a large number of groups.

B) Breaking data into many pieces.

C) Adding branches to a dendrogram.

D) Assigning the same datum to multiple clusters.

Answer: [A] When we oversegment, we're creating many clusters, possibly more than we need, in order to ensure that we have a sufficiently rich representation.

K-Means++ Which of the following is NOT TRUE of the K-Means++ algorithm?

A) It is a deterministic procedure.

B) It offers theoretical guarantees.

C) It is a way of initializing K-Means.

D) All three of these are true.

Answer: [A] K-Means++ is a randomized procedure.

Gap Statistic Which of the following is NOT TRUE of the Gap Statistic approach to selecting K in K-Means?

A It requires specifying a null model.

B It requires synthesizing reference data.

C It sums over the distances between clusters.

D All of these are true.

Answer: [C] It sums over the distances between data within the same cluster.

Lecture 3: Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering Which of the following is **not** true about hierarchical agglomerative clustering?

A) The number of clusters grows with the size of the data.

B) The method suffers from the curse of dimensionality.

C) The distance between two groups is determined solely by the closest pairs between the groups.

D) All of the above are true.

Answer: [C] The main decision in HAC is choosing the distance criterion. The shortest distance (single-linkage) is one of many distance metrics.

HAC Dendrogram True or False? In hierarchical agglomerative clustering, the distances between groups in the dendrogram produced by the clustering must be monotonically increasing for all linkage criteria: single-linkage, complete-linkage, average, and centroid.
Answer: False. The centroid criterion is not monotonic.

Linkage Criterion Which of the following linkage criterion in hierarchical agglomerative clustering produces compact clusters?

A) Complete-linkage

B) Average-linkage

C) Centroid-linkage

D) All of the above

Answer: [D] All of the above produce compact clusters. Single-linkage, however, produces stringy clusters (see Figure 4 in the notes).

Average-Linkage Criterion True or False? Average-linkage criterion produces the minimum spanning tree of the data.
Answer: False: Single-linkage criterion produces the MST.

Divisive Clustering True or False? Unlike K-means, divisive clustering is always deterministic.
Answer: False: Divisive clustering is non-deterministic.

Lecture 4: Principal Component Analysis

Principal Components True or false? In PCA, the top principal component is chosen using the smallest eigenvalue of the data covariance matrix.
Answer: False. The top principal component is chosen from the largest eigenvalue of the covariance matrix.

Principal Components True or false? In PCA, the top component is formed from a normalized non-linear combination of the features that has the largest variance.
Answer: False. Principal components are a linear combination of the features.

Uses of PCA Which of the following is NOT a use for principal component analysis?

A) Lossy data compression

B) Feature extraction

C) Data visualization

D) All of the above are true.

Answer: [D] All are uses of PCA.

PCA Interpretation Which of the following is a valid interpretation of PCA?

A) The principal components vectors give the directions in the feature space along which the data are the most variable.

B) The principal components provide low-dimensional surfaces that are closest to the data.

C) Both A and B

D) None of the above

Answer: [C] These are two different interpretations of PCA.

Principal Components True or false? The principal component vectors need to be unit length and orthogonal to each other.
Answer: True. The principal component vectors are orthonormal.

Lecture 5: Supervised Learning

Supervised Learning Which of the following is NOT an example of supervised learning?

A) Classification

B) Regression

C) Clustering

D) All of the above are examples of supervised learning.

Answer: [C] Clustering is an example of unsupervised learning.

Supervised Learning Which of the following statements is false?

A) One reason for pre-processing the dataset is to speed up computation.

B) In classification, we are interested in assigning each datum to a discrete-valued category.

C) Regularization is a technique used to deal with the problem of overfitting by penalizing complexity.

D) All of the above are true.

Answer: [D] All of the above are true.

Probability theory Which of the following statements is false?

A) The posterior distribution is proportional to the likelihood function times the prior.

B) The likelihood function is defined as the probability of the parameters given the data $p(w|D)$.

C) If X and Y are independent random variables, then $P(X, Y) = P(X)P(Y)$.

D) All of the above are true.

Answer: [B] The likelihood is defined as the probability of the data given the parameters $p(D|w)$.

Gaussian MLE Which of the following statements is true about the maximum likelihood estimator of a Gaussian distribution?

A) The maximum likelihood estimators are not available in closed form.

B) The maximum likelihood estimator for the mean is biased.

C) The maximum likelihood estimator for the variance is unbiased.

D) None of the above are true.

Answer: [D] None of the above are true.

Model Selection True or false? In S-fold cross-validation, the data is partitioned into S groups. Model scores are averaged over multiple runs in which the model is trained on one of the groups and tested on the remaining $S - 1$ groups of the data.
Answer: False. The $S - 1$ groups are used to train the model and its performance is evaluated on the remaining group.

Lecture 6: Regression

Combining Features True or false? Linear regression models the target values as a convex combination of the input features.
Answer: False. The weights are not required to sum to 1.

Linear Functions True or false? Linear regression requires the use of linear functions of the input variables.
Answer: False. More complex linear regression models include non-linear basis functions of the input variables.

Basis Functions Which of the following basis functions is closely related to the logistic function?

A) The identity basis

B) The polynomial basis

C) The tanh basis

D) The Fourier basis

Answer: [C] The hyperbolic tangent and logistic function are both sigmoids.

Ordinary Least Squares Which of the following is true about ordinary least squares (OLS) linear regression?

A) The OLS method requires that noise in the model must be Gaussian-distributed.

B) In a model with Gaussian noise, the parameters are the weights and the precision.

C) To estimate the parameter solutions, we can use the method of maximum likelihood.

D) Both A and B

E) Both B and C

Answer: [E] Both B and C.

Geometric Interpretation of Ordinary Least Squares Which of the following is true about the geometric interpretation of ordinary least squares (OLS) linear regression?

A) We can think of OLS as finding the point that corresponds to the choice of input that lies in the linear subspace closest to the target vector.

B) We can think of OLS as finding the point in R^d, where d is the number of dimensions of the data, that is equidistant from all the inputs.

C) Both A and B.

D) None of the above are true.

Answer: [A]

Lecture 7: Regression, Continued

Point Estimates True or false? A frequentist treatment of the regression problem involves making a point estimate of the weights w based on the data and tries to interpret the uncertainty of the estimate by averaging over an ensemble of different data sets.
Answer: True.

Mean Squared Error True or false? The expected squared difference between the prediction function and the regression function is a function of the squared bias of the two functions and the variance of the prediction function.
Answer: True.

Bayesian Linear Regression Which of the following is **true** regarding a Bayesian approach to linear regression?

A) A Bayesian modeling approach places a prior probability distribution over the weights, rather than using a point estimate, which is prone to overfitting.

B) When the likelihood function is a Gaussian with known precision, the conjugate prior for the weights is a Gaussian distribution.

C) The posterior distribution can be computed sequentially by setting the prior distribution to the posterior from the previous step.

D) Both A and B.

E) Both A and C.

F) A, B, and C are true.

Answer: [F] A, B, and C are all true.

MAP Estimate True or false? The MAP estimate is the mode of the prior distribution.
Answer: False. The MAP estimate is the mode of the posterior distribution.

Predictive distribution Which of the following is FALSE about the predictive distribution of a Bayesian linear regression model?

A) The predictive distribution gives a distribution on target values for future unseen data.

B) The predictive distribution involves integrating out the uncertainty associated with the parameter weights in the posterior distribution.

C) The predictive distribution of a Gaussian posterior with unknown weights and precision is Gaussian.

D) All of the above are true.

Answer: [C] When the precision is known, the predictive distribution is Gaussian. However, when the precision is also unknown, the predictive distribution takes the form of a Student's t-distribution.

Lecture 8: Model Selection

Avoiding Overfitting Which of the following are methods to avoid overfitting?

A) Regularizing, using cross-validation to set the parameters

B) Finding a distribution over model parameters instead of finding a point estimate

C) Adding more parameters to the model

D) Both A and B.

E) Both A and C.

F) A, B, and C are true.

Answer: [D] Both A and B. Adding parameters to the model increases the chance of overfitting.

Model Selection True or false? Model selection refers to optimizing the values taken on by parameters in a particular model.
Answer: False. Model selection refers to deciding which class of model to use.

Marginal Likelihood True or false? A common method of Bayesian model selection is calculating the marginal likelihood.
Answer: True.

Kullback-Leibler Divergence True or False? The KL divergence of two probability distributions goes to ∞ when they are equal.
Answer: False. It goes to 0.

Marginal Likelihood The marginal likelihood penalizes complexity because:

A) It explicitly adds a term that penalizes the number of parameters in the model.

B) It leads to a tradeoff between the ability to fit a broader set of possible data and the goodness of fit.

C) All of the above.

D) None of the above.

Answer: [B] Marginal likelihood penalizes complexity because even though more complex models fit the data better, they are also represent a broader set of hypotheses that can fit a larger set of possible data.

Lecture 9: Linear Classification

Linear Separability True or false? A dataset is said to be linearly separable if points in different classes *cannot* be divided into two classes by any linear equation.
Answer: False.

Discriminant Function True or false? If our discriminant function is $y(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x} + w_0$, why is w_0 is called the bias term.
Answer: True.

Multi-class Linear Discriminants A one-versus the rest classifier:

1. Uses a single classifier that maps a data point to one of K classes.
2. Can lead to ambiguity in classification.
3. Uses many classifiers, each of which indicates whether a point is in a particular class.
4. A and B
5. B and C

Answer: [E] Since many classifiers are being used, there could be points that are marked to not be in any of the classes.

Fisher's Linear Discriminant The Fisher's Linear Discriminant:

1. Projects the data in a way that maximizes variances within each class.
2. Projects the data in a way that maximizes separation between each classes.
3. All of the above.
4. None of the above.

Answer: [B] Definition of Fisher's Linear Discriminant, see Bishop, p. 187.

Perceptrons A perceptron

A) Typically gives continuous output.

B) Typically gives output of -1 or 1.

C) Uses a linear activation function to get the output.

D) None of the above.

Answer: [B] A perceptron's activation function gives output of -1 or 1.

Lecture 10: Probabilistic Classification

Probabilistic Generative Models True or False? Probabilistic generative models never give linear decision boundaries.
Answer: False. Some simple generative models give linear decision boundaries, such as Gaussians with identical variances.

Generative Models A probabilistic generative model

A) is in contrast with discriminative models.

B) views data as generated by sampling from several classes, and then sampling a class-conditional value.

C) cannot be used to calculate the posterior explicitly.

D) A and B

E) B and C

Answer: [D] Generative models place full distributions on the data and are taken to be different than discriminative models, which only model the conditional distribution of the label. They model the probability of the class and the probability of the features given the class.

Exponential Family Which of the following is in the exponential family of distributions?

A) Binomial distribution

B) Gaussian distribution

C) Both A and B

D) All of the above

Answer: [D] These are both in the exponential family, as seen in Bishop, Section 4.2.4.

Logistic Regression True or false? Like regression, logistic regression takes in continuous inputs and returns estimates of the outputs.
Answer: False. Logistic regression is a binary classifier that takes in continuous inputs and returns the probability of being in each class.

Probit Regression Probit regression uses the following link function:

A) Identity function

B) Sigmoid function

C) Cumulative distribution of Gaussian

D) Any of the above

Answer: [C] The probit function is the cumulative distribution function of a zero mean, unit variance Gaussian.

Lecture 11: Neural Networks

Neural Network Speed True or False? As compared to other models with similar goals, neural networks are faster to train but slower to evaluate.
Answer: False. Neural networks are typically very slow to train, but can be evaluated on test data very rapidly, as they are feed-forward.

Layer-to-layer transformations How are the outputs from one layer typically transformed to inputs for the next layer in feed-forward neural networks?

A) Inputs for one layer are linear combinations of outputs from the previous layer.

B) Inputs for one layer are found by taking linear combinations of outputs from the previous layer, and then applying a non-continuous activation function.

C) Inputs for one layer are found by taking linear combinations of outputs from the previous layer, and then applying a continuous, but non-differentiable activation function.

D) Inputs for one layer are found by taking linear combinations of outputs from the previous layer, and then applying a continuous and differentiable activation function.

E) None of the above

Answer: [D] See Bishop, Page 227.

Activation Functions What are common activation functions for multi-layer neural networks?

A) Hyperbolic tangent function

B) Logistic function

C) All of the above.

D) None of the above.

Answer: [C] S-shaped sigmoid curves are very popular.

Gradient Descent True or false? Stochastic gradient descent is often to handle gradient descent for large streams of data points.
Answer: True. It is an online version of gradient descent.

Backpropagation Which of the following is true of back-propagation?

A) It is a way to perform gradient descent.

B) The gradient for a particular weight in the network can be calculated directly from error terms of nodes for lower layers.

C) Since the error function is convex, we only need to run back-propagation for one iteration.

D) A and B

E) A and C

Answer: [A] The gradient for a particular weight is calculated from error terms of nodes for layers after this weight. Also, back-propagation normally needs to be run for many iterations.

Lecture 13: Decision Trees

Decision Trees and Interpretability True or False? Decision trees, as compared to other models with similar goals, are hard to interpret.
Answer: False. Decision trees are easier to interpret since they have decision nodes that represent features.

Learning from Truth Tables True or False? Looking at all possible truth tables is a computationally feasible way to learn a decision tree.
Answer: False, there are 2^{2^D} possible truth tables if D is the number of features.

Entropy If we are more certain about an event, the Shannon entropy gets:

A) Bigger.

B) Smaller.

C) Depends.

D) Stays the same.

Answer: [B] The Shannon entropy gets smaller as we get more certain since we need fewer bits to represent certainty. We can also see this through the formula in the notes.

Mutual Information and Conditional Entropy Do we choose attributes based on maximizing mutual information or conditional entropy when training a decision tree? Why?

A) Conditional entropy, because it describes how much uncertainty we have in the outcome after seeing a given feature.

B) Mutual information, because it describes how much uncertainty we have in the outcome after seeing a given feature.

C) Mutual information, because it describes the *additional* information about the outcome we get from a given feature

D) Mutual Information, because it tells us the extent to which a feature and the outcome are linearly related.

Answer: [C] See course notes, Section 1.4.

Regularizing Decision Trees Which of the following are ways to regularize when training decision trees?

A) Early stopping.

B) Pre-pruning or post-pruning.

C) Adding nodes only if their mutual information is low enough.

D) A and B

E) All of the above.

Answer: [D] Early stopping and pre-pruning are ways to regularize. Another way of regularization is to add nodes only if their mutual information is high enough.