Question 1

We discussed how clustering can be used to help address the curse of dimensionality

○ [✓ True]

○ False

Question 2

The K-means algorithm will generate a clustering that yields the minimum sum-of-squared error (SSE).

○ True

○ [✓ False]

Question 3

The K-Means algorithm uses which type of cluster? Select the **one** best answer.

○ Contiguity-based

○ [✓ Center-based clusters]

○ Well-Separated

○ Density-based clusters

Question 4

A dendogram is most often used to represent a partitional clustering.

○ True

○ [✓ False]

Question 6

Which one of the following statements about the ensemble methods that vary the training data to form multiple classifiers is correct?

○ Bagging and Boosting vary the training data but Random Forest does not.

○ Random Forest varies the training data but Bagging and Boosting do not.

○ None of Bagging, Boosting, and Random Forest vary the training data.

○ [✓ Bagging, Boosting, and Random Forest vary the training data.]

Question 7

Generally speaking, what is the key to an ensemble of classifiers doing better than a single classifier? (Hint: the answer has to do with the *relationship* between the classifiers and not the accuracy of the base classifiers).

Independence of each classifier

Question 8

The bagging ensemble method can be used with all classification algorithms.

○ [✓ True]

○ False

Question 10

One approach to dealing with class imbalance is to ignore the majority class examples and learn only from the minority class examples.

○ [✓ True]

○ False

Question 11

In the context of medical diagnosis, what is the most common relationship between the cost of False Positives (FP) and False Negatives (FN). Select the **one** best answer.

○ Cost of FP = 0.

○ Cost of FP = Cost of FN.

○ [✓ Cost of FN > Cost of FP.]

○ Cost of FP > Cost of FN.

Question 12

If a text document is represented using the "bag of words" approach, then the word ordering information will be preserved.

○ True

○ [✓ False]

Question 13

The Simple Matching Coefficient (SMC) metric is more appropriate than the Jaccard Coefficient when determining the similarity between sparse binary vectors.

○ True

○ [✓ False]

Question 14

As you investigate your data in preparation for building a nearest-neighbor classification model, you find that the correlation between two features, f1 and f4, is 1.0. What action should you take based on this information?

The features contain the same information so you should remove one since it is redundant.

Question 15

Association rules imply causality (i.e., if A -> B then A causes B to occur).

○ True

○ [✓ False]

Question 16

Which of the following implications/consequences of the Apriori property **are true**. *Select all that apply*.

□ If an itemset is frequent, then all of its supersets must be frequent.

□ [✓ If an itemset is frequent, then all of its subsets must be frequent.]

□ [✓ If an itemset is *not* frequent, then all of its supersets much not be frequent.]

□ If an itemset is *not* frequent, then all of its subsets much not be frequent.

Question 17

In association rule mining, the quantity of an item in a transaction does not matter- it does not matter if milk is purchased one time or two times in a given transaction.

○ [✓ True]

○ False

Question 18

Entropy measures randomness (impurity). Given two class values, "+" and "-", which class probabilities yield the maximum entropy value?

○ P(+) = 0.75 and P(-) = 0.25

○ P(+) = 0 and P(-) = 1

○ P(+) = 1 and P(-) = 0

○ [✓ P(+) = 0.5 and P(-) = 0.5]