# FODS: Exam Review

## Part A (True/False)

All questions should be answered by indicating whether the statements are True (T) or False (F). Please indicate this next to each statement and not anywhere else.

1. In the CRISP-DM process, the Data Understanding phase comes after the Data Preparation phase.
2. An association rule with low confidence can have high support in a dataset.
3. PCA reduces dimensionality by transforming variables into a smaller set of uncorrelated components, which are linear combinations of the original variables.
4. The principal components in PCA always explain 100% of the variance in the original dataset.
5. K-means and K-medoids have the same computational complexity in terms of time and space.
6. The silhouette score is only suitable for evaluating the performance of the k-means clustering algorithm.
7. Lasso Regression (L1) can lead to coefficients with zero value, as some features can be completely neglected for the evaluation of the final output.
8. A decision tree will never overfit the training data, regardless of its depth and complexity.
9. The decision boundary of 1-NN is always linear.
10. Perceptron is a type of unsupervised learning algorithm.
11. Cross-validation guarantees that a model will perform well on unseen data.
12. Accuracy is a suitable evaluation metric for imbalanced datasets.
13. Exploitation is the primary focus in reinforcement learning and should always be prioritized over exploration.

## Part B (Multiple Choice)

**Question 1: Which of the following best describes the purpose of the CRISP-DM methodology?**

(a) To provide a statistical analysis framework.

(b) To outline a standard process for data mining projects.

(c) To increase computational efficiency in data analysis.

(d) To eliminate the need for data preprocessing.

**Question 2: In the context of cross-validation, what does k-fold cross-validation entail?**

(a) Splitting the data into k parts and training k different models on the same data.

(b) Dividing the data into k subsets and using k-1 for training and 1 for testing iteratively.

(c) Using a single train/test split for model evaluation.

(d) Training the model only once and testing it on the entire dataset.

**Question 3: What is the primary goal of Principal Component Analysis (PCA)?**

(a) To maximize the number of features in a dataset.

(b) To transform the data into a higher-dimensional space.

(c) To reduce the dimensionality while retaining as much variance as possible.

(d) To normalize the data.

**Question 4: What is a potential disadvantage of using K-Means clustering?**

(a) It is computationally expensive.

(b) It requires the number of clusters to be defined beforehand.

(c) It cannot be used for large datasets.

(d) It always produces the same clustering results regardless of initialization.

**Question 5: Which activation function is commonly used in the output layer of a binary classification neural network?**

(a) Softmax

(b) Sigmoid

(c) ReLU

(d) Linear

# Part C (Open Questions)

### Question 1: Frequent Itemset Mining

Consider the following transaction database $D_1$. Each transaction contains a set of items:

| Transaction ID | Items |
|---|---|
| 1 | {A, B, C} |
| 2 | {B, D} |
| 3 | {A, B} |
| 4 | {B, C, D} |
| 5 | {B, C} |
| 6 | {A, B, C} |
| 7 | {A, C, D} |
| 8 | {A, C} |

- Calculate and report the absolute support for every possible itemset that can be generated using the universe I = {A, B, C, D}, based on the transactions in the database. Identify and report the frequent itemsets based on a minimum absolute support threshold of 2.
- Among the frequent itemsets, identify and report the closed itemsets. Explain in one sentence why you chose them.
- Among the frequent itemsets, identify and report the maximal itemsets. Explain in one sentence why you chose them.
- Considering the frequent itemsets that you have identified above, list any association rules with 2 items, e.g., A → B, with a confidence of at least 70% that can be derived from the frequent itemsets and explain why you chose them.

*Recall the definition of confidence:* confidence (X→Y) = supp (X, Y) / supp (X)

### Question 2: Dimensionality Reduction

- Explain briefly what is the primary goal of both Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS) in terms of dimensionality reduction.
- Describe briefly the steps used by PCA and compare them with the approach used by MDS.
- Explain briefly which method, PCA or MDS, you would prefer if you are primarily concerned with preserving the distances between points in your high-dimensional dataset and why.

### Question 3: Clustering

- Compare and contrast K-Means clustering with hierarchical clustering. Discuss the advantages and disadvantages of each approach and when one might be preferred over the other.
- Discuss how you would use the Silhouette coefficient for determining the value of parameter k for the k-means algorithm.

- According to the curse of dimensionality, as the number of data dimensions increases, the notion of similarity becomes meaningless. Explain the rationale behind this and discuss at least one way to go about solving the problem.

### Question 4: Classification

Consider a set of five training examples given as $((x_i, y_i), c_i))$ values, where $x_i$ and $y_i$ are the two attribute values (positive integers) and $c_i$ is the binary class label: $((1,1) - 1), ((1,7) + 1), ((3,3) + 1), ((5,4) - 1), ((2,5) - 1)$. Classify a test example at coordinates (3, 6) using a k-NN classifier with $k = 3$ and Manhattan distance as follows: (Your answer should be either +1 or -1. You should write your complete answer, not just a single answer like -1 or +1)

$$d((u, v), (p, q)) = |u - p| + |v - q|$$

### Question 5: Model Evaluation

You are analyzing the performance of a binary classification model used to identify rare and life-threatening medical conditions. The confusion matrix for this model is as follows:

| | Predicted positive | Predicted negative |
|---|---|---|
| Actual positive | 42 | 8 |
| Actual negative | 13 | 9379 |

- Discuss the potential consequences of misclassification in this medical context. In particular, address the practical interpretation of false positives and false negatives.
- Describe a scenario in which maximizing precision might be more important than maximizing recall, and another scenario where the opposite is true. Discuss which scenario is more suited here.

### Question 6: Deep Learning

Imagine you are tasked with building a basic neural network to classify whether an email is spam or not (binary classification).

- Explain briefly what the input and output of the neural network would look like for this task.
- Explain briefly which activation function you would use in the output layer and why.
- If you were working on a multi-class classification problem instead of binary, explain briefly which activation function you would use instead.