CS5228: Knowledge Discovery & Data Mining

Midterm Exam

Date: Friday, 4 October 2024, Time: 19:00–20:00

Submission Instructions

1. Please read **ALL** instructions carefully.
2. All the assessment is be done using Examplify; the assessment contains
   (a) MCQ/MRQ: Questions 1–16 (30 points)
   (b) Essay: Question 17–21 (30 points)
3. The total number of points is 60
4. This is an open-book assessment.
5. Your Internet connection will be blocked for the duration of the assessment.
6. This assessment starts at 19:00 and ends at 20:00.
   - Submit your answers by 20:00.
   - No additional time will be given to submit.
7. For the MCQ/MRQ questions: Please note that the order of the answers might differ between this PDF and Examplify due to randomization!
8. Failure to follow each of the instructions above may result in deduction of your marks.

Good Luck!

MCQ/MRQ

Q1: (2 points) (MRQ) Given is a dataset with only numerical features. Which operation(s) on the dataset will **NOT** affect the result of DBSCAN when using the same parameters for ϵ and *MinPts* and the Euclidean Distance metric? (Note: Ignore the corner case where a point can be border point of different clusters!)

- ◯ Standardization of the data
- ◯ Multiplying all feature values by the same constant
- √ **Adding the same constant to all feature values**
- √ **Re-ordering the samples in the dataset**
- ◯ None of the above

Q2: (1 point) (MCQ) What is the reason for the **Silhouette Coefficient** penalizing a large number of clusters?

- ◯ The cohesion goes up
- ◯ The cohesion goes down
- ◯ The separation goes up
- √ **The separation goes down**
- ◯ None of the above

Q3: (2 points) (MRQ) Which of the following statements about Apriori Algorithm for finding Frequent Itemsets are always **TRUE**?

- √ **If an itemset X is frequent, *all* of the subsets of X are frequent**
- ◯ If an itemset X is infrequent, *none* of the subsets of X are frequent
- ◯ If an itemset X is frequent, *some* of the supersets of X are frequent
- √ **If an itemset X is infrequent, *any* of the supersets of X are infrequent**
- ◯ None of the above

Q4: (2 points) (MRQ) Which factors will **NOT** affect the runtime performance of the Apriori Algorithm for finding Frequent Itemsets?

- ◯ The minimum support threshold
- ◯ The size of the dataset
- √ **The minimum confidence threshold**
- ◯ The number of frequent itemsets in the dataset
- ◯ None of the above

Q5: (2 points) (MRQ) Which of the following statements about the relationship between support and confidence for finding association rules are **TRUE**?

- √ **The support of a rule is never larger than its confidence**
- ◯ Confidence is used to prune itemsets before calculating their support
- ◯ A high support guarantees a high confidence
- √ **Rules with low support but high confidence can still be useful**
- ◯ None of the above

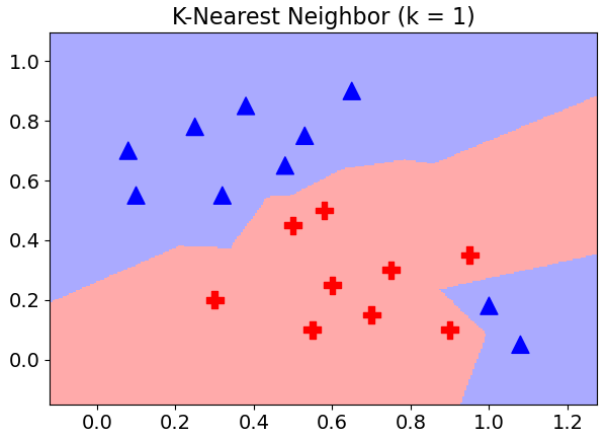
Q6: (2 points) (MCQ) Assume a transaction database containing 5 distinct items. 2 of those items are infrequent and 3 items are frequent. What is **the minimum and maximum number of frequent itemsets** in the database? The answers below are given in the form (minimum, maximum).

- ◯ (0, 8)
- ◯ (3, 8)
- ◯ (3, 11)
- √ **(3, 7)**
- ◯ None of the above

Q7: (1 point) (MRQ) Assume you want to build a system for predicting high-value investment opportunities for your clients. Assuming that a high-value investment is your positive class (or Class 1), what are your goals so your clients are unlikely to lose their money?

- ◯ Maximize recall
- √ **Maximize precision**
- √ **Minimize number of false positives**
- ◯ Minimize number of false negatives
- ◯ None of the above

Q8: (2 points) (MCQ) As an example, the figure below shows the simple binary classification dataset, including the decision boundaries after running a K-Nearest Neighbor (KNN) classifier with $K = 1$.



For this dataset and $K = 1$ there are **3 separate regions** indicating how a unseen data point would be labeled if it would fall into any of those regions.

Now assume a classification dataset with **18 data points** and **3 classes**: 5 Red, 6 Blue, and 7 Green. For any arbitrary distribution of the data points and any arbitrary choice of K, what is **the minimum and maximum possiblenumber of separate regions** after training a KNN classifier? The answers below are given in the form (minimum, maximum).

- ◯ $(0, 3)$
- ◯ $(1, 3)$
- ◯ $(1, 7)$
- √ $(1, 18)$
- ◯ None of the above

Q9: (2 points) (MRQ) Which statements about the difference between a K-Nearest Neighbor (KNN) classifier and a Decision Tree classifier are **TRUE**?

- ◯ Only Decision Trees support both categorical and numerical features
- ◯ Only Decision Trees result in piecewise linear decision boundaries
- ◯ Only Decision Trees can result in overfitting models
- √ **Only Decision Trees are not affected by feature scaling**
- ◯ None of the above

Q10: (2 points) (MRQ) Which of the following statements about increasing the maximum allowed depth of a Decision Tree classifier are **TRUE**?

- √ **The risk of overfitting tends to increase**
- ◯ The risk of underfitting tends to increase
- √ **The training error tends to go down**
- √ **The decision boundaries tend to get more complex**
- ◯ None of the above

Q11: (2 points) (MRQ) Which of the following are methods Random Forests use to add randomness and increase model diversity?

- √ **Randomly selecting subsets of features for splitting each node**
- √ **Randomly selecting subsets of data through bootstrap sampling**
- ◯ Randomly initializing weights in each tree
- ◯ Randomly reducing the maximum depth of some trees
- ◯ None of the above

Q12: (2 points) (MRQ) Which of the following statements reflect potential drawbacks or limitations of Random Forests?

- √ **It can be computationally expensive with large datasets**
- ◯ It is prone to overfitting with an increasing number of trees
- √ **It can be difficult to interpret compared to a single Decision Tree**
- ◯ It is not suitable for high-dimensional datasets
- ◯ None of the above

Q13: (2 points) (MRQ) Which of the following describe the role of a weak learner in AdaBoost?

- √ **A weak learner performs slightly better than random guessing**
- ◯ A weak learner is trained independently from other learners
- ◯ A weak learner is one that makes highly accurate predictions
- √ **A weak learner contributes to the final prediction through weighted voting**
- ◯ None of the above

Q14: (2 points) (MRQ) Which of the following statements best describe the boosting process in Gradient Boosted Trees?

- ◯ Predictions from all trees are averaged to make the final prediction
- √ **It is prone to overfitting with an increasing depth of the trees**
- √ **Each tree is built to correct the errors of the previous tree**
- √ **All trees are trained sequentially**
- ◯ None of the above

Q15: (2 points) (MRQ) Which of the following methods help prevent overfitting in Gradient Boosted Trees?

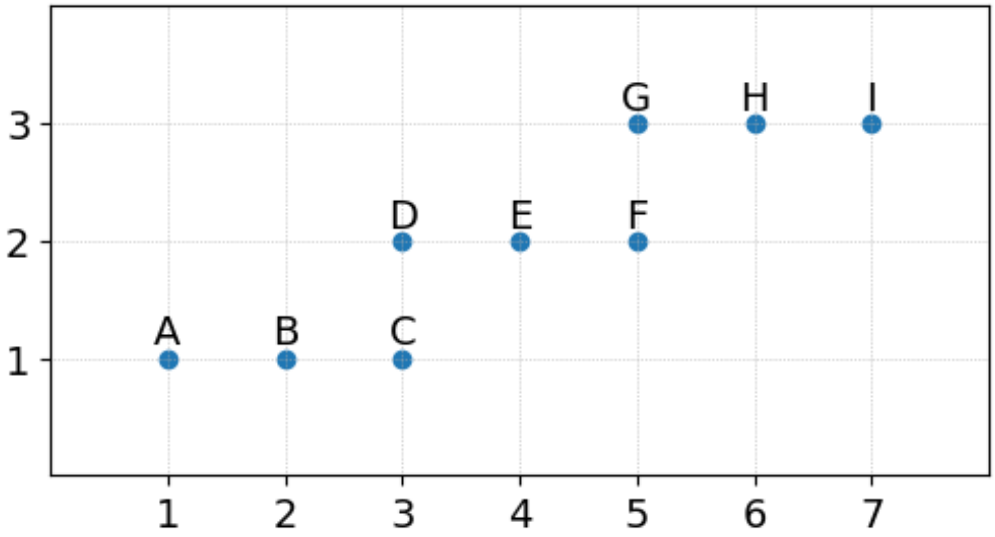
- √ **Reducing the learning rate**
- ◯ Increasing the number of trees
- √ **Using early stopping**
- √ **Limiting the depth of the trees**
- ◯ None of the above

Q16: (2 points) (MRQ) Which of the following are key differences between Gradient Boosting and Random Forests?

- √ **Gradient Boosting builds trees sequentially, while Random Forests build trees in parallel**
- √ **Gradient Boosting focuses on correcting the errors of previous trees, while Random Forests average predictions from many independent trees**
- ◯ Random Forests are more prone to overfitting than Gradient Boosting
- √ **Gradient Boosting requires tuning a learning rate, while Random Forests do not**
- ◯ None of the above

Clustering

Q17: (7 points) Given is the following simple dataset containing the nine points $A, B, ..., I$ in the Euclidean space:



Your task is to find **three** clusters with AGNES. For deterministic outcomes, should there be any two pairs of clusters eligible for merging with the same inter-cluster distance, i.e., $d(C_1, C_2) = d(C_3, C_4)$, the pair with the cluster nearest to coordinate (0, 0) should be prioritized and merged first. Answer the following questions:

(a) Perform AGNES with **Single Linkage**. List the sequence of merges and the resulting three clusters. For example, if a cluster with points X and Y is merged with a cluster with point Z, you should write *XY-Z*. (3 points)

(b) Perform AGNES with **Complete Linkage**. List the sequence of merges and the resulting three clusters below. For example, if a cluster with points X and Y is merged with a cluster with point Z, you should write *XY-Z*. (3 points)

(c) Which of the two linkage methods results in the better clustering result and why? Briefly explain your answer. (1 point)

Solution:

1.
   - Merges: A-B, AB-C, ABC-D, ABCD-E, ABCDE-F, ABCDEF-G
   - Clusters: ABCDEFG, H, I
2.
   - Merges: A-B, C-D, E-F, G-H, GH-I, AB-CD
   - Clusters: ABCD, EF, GHI
3. Complete linkage had the better clustering result. Complete linkage is less susceptible to noise and thus, "chaining".

Association Rule Mining

Q18: (4 points) Assume a small transaction dataset with 4 different items A, B, C, and D. Also assume the Apriori Algorithm identified the following five 2-itemsets that satisfy a user given support threshold, i.e., these five 2-itemsets are frequent itemsets:

$$\{A, B\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}$$

What initial candidate 3-itemsets are created by the Apriori Algorithm algorithm; which of those survive subset pruning?

Solution: Initial candidate 3-itemsets:

$$\{A, B, C\}, \{A, C, D\}, \{A, B, D\}, \{B, C, D\}$$

Remaining Itemsets after pruning:

$$\{A, B, D\}, \{B, C, D\}$$

Q19: (5 points) Assume you have to mine association rules for a very large transaction database which contains 10,000,000 transactions. How could sampling be used to speed up association rule mining (ARM)? Describe an algorithm which uses sampling to speed up ARM but is still likely to return the same rules as running ARM on the original dataset!

Solution: One possible solution

- Run the Association Rule Mining algorithm for a much smaller sample (e.g. 500,000 transactions) with a **(slightly) lower** support and confidence threshold obtaining a set of association rules R
- Only for those rules in R, go through the complete transaction database and compute their "true" support and confidence value; prune all rules which violate the confidence or support thresholds.
- Return the surviving rules.

Evaluation of Classifiers

Q20: (6 points) Given is the following confusion matrix for a binary classifier which is missing the ground truth values for Class 0:

		Ground Truth	
		Class 1	Class 0
Prediction	Class 1	$TP = 90$	$FP =???$
	Class 0	$FN = 30$	$TN =???$

However, we also know the values for the following evaluation metrics:

- Accuracy $= 0.8 = 4/5$
- Precision $= 0.9 = 9/10$
- Recall $= 0.75 = 3/4$

Compute the missing values for the number of **False Positives (FP)** and the number of **True Negatives (TN)**!

Solution: First, compute FP using precision:

$$Precision = \frac{TP}{TP + FP} = 9/10$$

$$FP = \frac{10 \cdot TP}{9} - TP = \frac{10 \cdot 90}{9} - 90 = 10$$

Now we can use the accuracy to compute TN:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{90 + TN}{130 + TN} = 4/5$$

$$450 + 5TN = 520 + 4TN$$

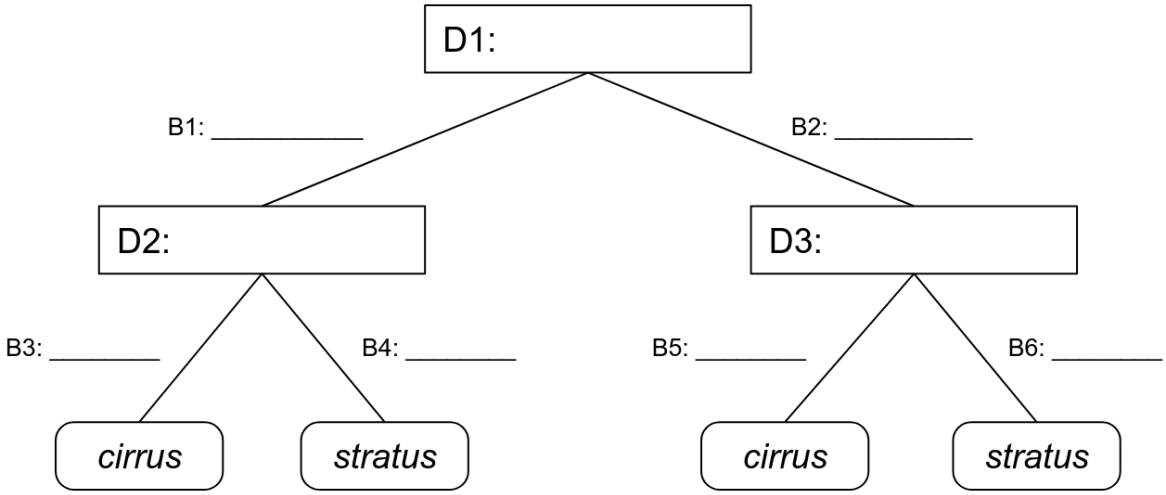
$$TN = 70$$

Classifiers

Q21: (8 points) Assume the following dataset for training a classifier to predict if a cloud in the sky is a *cirrus cloud* or *stratus cloud* based on its color, altitude, and size:

#	**Color**	**Altitude**	**Size**	**Class**
1	light	high	small	cirrus
2	dark	high	small	cirrus
3	light	high	small	cirrus
4	dark	low	large	cirrus
5	dark	high	large	stratus
6	light	low	small	stratus
7	dark	low	small	stratus
8	light	low	small	stratus

A Decision Tree trained on this dataset using the features **Color**, **Altitude**, and **Size** to perfectly predict the **Class** (*cirrus* or *stratus*) will have the following structure:

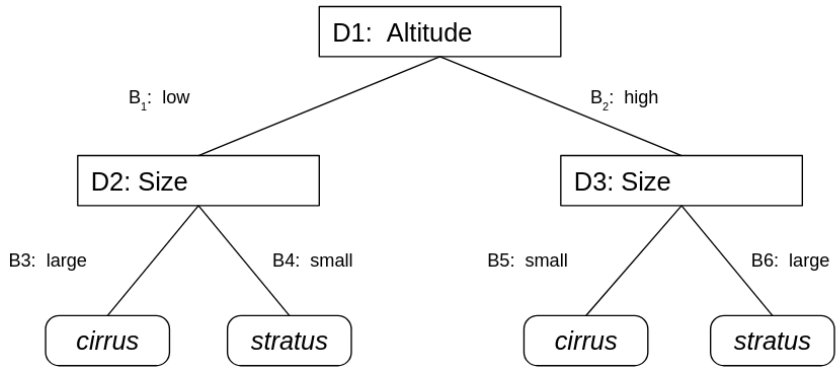


(a) Complete the Decision Tree above by providing all the missing labels (6 points):

- D1-3: the features used for the respective splits
- B1-6: the feature values specifying the branch

Hint: The dataset is very simple, so there should be no need for any calculations; all features are also binary, i.e., having only 2 different values each.

Solution:



Note: The branches for the root split can be mirrored without changing the correctness.

(b) In general, the same feature might be used multiple times to split a node along the same path from the root node to a leaf node in a Decision Tree. This cannot happen for this dataset. Briefly explain why! (2 points)

Solution:

- All attributes/features have only two different values each (i.e., binary features)
- Whatever feature is chosen, say "Size" with the two values "small" and "large", the "small" samples will be in one child node and the "large" samples will be in the other child node.
- Since now all samples on a child node have the same value for "Size", this feature no longer provides and discriminatory power.
- In fact, binary features can only ever used only once along a path