Data Mining Algorithms I: Final Exam

Lecturer: Prof. Seidl

25.02.2015

The following document is a transcript from memory. There is no guarantee for correctness or completeness.

- 70 points in total
- 10 exercises

1)

(a)

Steps of KDD

(b)

Give the definition of *Classification*.

(c)

Name an evalution measure for *Clustering* and *Classification*.

(d)

Name examples for distributive, algebraic, holistic measure.

2) Frequent Itemset Mining

(a)

Given set of frequent itemsets L. Is L result of Apriori-Agorithm?

(b)

Prove or disprove: If all non-empty subsets are frequent, then the itemset itself is frequent.

(c)

Give an example for $conf(A \Rightarrow B) \geq 60\%$ and $corr_{A,B} < 1$.

(d)

Given an FP-tree, complete the FP-growth algorithm.

3)

(a)

Given a figure of data points. Specify medoid, mode and median by drawing.

(b)

Given clustering. Is it a valid k-means clustering using Manhattan distance and $k = 2$.

(c)

Describe the E and M step of EM-algorithms. What is the major difference between the E-step and the according step in k-means?

4) Agglomerative Hierarchical Clustering

(a)

Give the definition of complete link

(b)

Given a set of data points, draw dendrogram using single link and Manhattan Distance.

5) Bayes

Query $q = (q_1, \ldots, q_d)$.

(a)

Decision Rule for Bayes classifier.

(b)

Decision Rule for naive Bayes classifier.

(c)

Consider data set of 16 black circles, 9 grey triangles in $\mathbb{R}^2$, as shown in Figure together with their marginal distribution

(i)

Classify the query q (depicted) with naive Bayes.

(ii)

Classify the query q (depicted) with (non-naive) Bayes.

6) Z-Order

Z-Order of level 2 of $[0, \ldots, 7] \times [0, \ldots, 7]$.

(a)

What are the Z-values of (five points, all in $[0, 3] \times [0, 3]$)

1.

(b)

Assume the values in a) were ... Insert them into the given B-Tree. Draw after each split.

(c)

Delete the following entries (4) from the following B-Tree. Draw after each deletion.

7) Regression Trees

Consider data points $P_1 = (0, 1), P_2 = (1, 3), P_3 = (2, 3), P_4 = (3, 4), P_5 = (4, 4), P_6 = (5, 3), P_7 = (6, 2)$ and the split $T_1 = \{P_1, \ldots, P_4\}, T_2 = \{P_5, \ldots, P_7\}$.

(a)

Compute the regression functions. Hint: $(\left(\tilde{X}^T\tilde{X}\right)^{-1}$ given for both parts).

(b)

Let $\text{imp}(T) = 0.9$. Using the variance of the residuals, is the split significant using the impurity ratio with $\tau = 0.5$?

8) DBSCAN

(a)

Given a dataset and $\epsilon = 2, MinPts = 5$. Is the given clustering a valid output of DBSCAN?

(b)

Name two difference between the output of DBSCAN and k-means

(c)

Assume ϵ is set reasonable. What influence does the choice of *MinPts* has on the result?

9) Decision Tree learning

(a)

Given an incomplete decision tree, complete it using the *Gini-index*.

(b)

What exactly is overfitting in a decision tree. How can this be avoided?

10) ???