Online Test 1: Fundamentals

Test und Assessment – Druckansicht

Online Test 1: Fundamentals

Datum: Sun Mar 13 14:10:44 2022 Maximale Punktezahl: 16

Frage 1 - KDD Process (1 Punkt) [ID: 1059015]

Order the steps and data in the KDD process! Hint: Steps and data should be alternating.

[Data] Data Sources

[Step] Integration & Cleaning

[Data] Data Warehouse

[Step] Selection

[Data] Task-Relevant Data

[Step] Transformation

[Data] Transformed Data

[Step] Mining

[Data] Patterns

[Step] Evaluation / Interpretation

[Data] Knowledge

See the lecture.

Frage 2 - Supervised vs Unsupervised (1 Punkt) [ID: 1051772]

Match the different categories of data-mining approaches with the paradigms "supervised" and "unsupervised"!

Supervised	passt zu	Classification	*(0.2 Punkte)*
Supervised	passt zu	Regression	*(0.2 Punkte)*
Unsupervised	passt zu	Association Rule Mining	*(0.2 Punkte)*
Unsupervised	passt zu	Clustering	*(0.2 Punkte)*
Unsupervised	passt zu	Outlier Detection	*(0.2 Punkte)*

Classification and regression use a target variable (categorical or numeric), approaches from the other three categories usually don't. However, this also depends on the concrete approach and the available data. For example, there also are supervised and semi-supervised outlier-detection approaches.

Frage 3 - One Rule (1 Punkt) [ID: 1059018]

Assume we want to analyze which factors led to passing or failing the "Data Science 1" exam. We have the following frequency tables:

Python_knowledge	**Failed**	**Passed**
No	4	16
Basic	5	25
Intermediate	10	30
Pro	1	9

Loves_data_science	**Failed**	**Passed**
Not at all	10	9
A bit	9	57
More than anything else in the world	1	14

Attended_exercises	**Failed**	**Passed**
None	13	10
Some	7	50
All	0	20

The feature chosen by One Rule is [Attended_exercises] *(0.5 Punkte)* .

The corresponding total error (as a fraction) is [0.17] *(0.5 Punkte)* .

For each feature value, we look which class is the least frequent and sum up those frequencies to get the error per feature. We choose the feature with the lowest error.

Error for **Python_knowledge**: (4 + 5 + 10 + 1) / 100 = 0.2

Error for **Loves_data_science**: (9 + 9 + 1) / 100 = 0.19

Error for **Attended_exercises**: (10 + 7 + 0) / 100 = 0.17

Note that the data is quite imbalanced. In most cases, predicting the majority class "Passed" is best. For **Python_knowledge**, the classifier only reproduces this baseline.

Frage 4 - Categories of Data (1 Punkt) [ID: 1051777]

Match the types of data with exemplary features!

nominal	passt zu	"Color", range of values = {green, blue, red, yellow, ...}	*(0.25 Punkte)*
ordinal	passt zu	"Grade", range of values = {outstanding, good, satisfactory, barely passing, failed}	*(0.25 Punkte)*
discrete	passt zu	"Dice Roll", range of values = {1, 2, 3, 4, 5, 6}	*(0.25 Punkte)*

conti-nuous	passt zu	"Account Balance", range of values = real numbers	*(0.25 Punkte)*

"grade" can, of course, also be coded discretely in a numerical way. The boundary between discrete and continuous might be blurry. The lecture draws the line between a finite amount of values (-> discrete) and an infinite amount of values (-> continuous). In memory, of course, we always store somewhat discrete values.

Frage 5 - Aggregation Functions (1 Punkt) [ID: 1051773]

Which of the following aggregation functions are distributive?

- [x] Minimum *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*
- [ ] Mean (= average) *(Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.2 Punkte)*
- [ ] Median *(Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.2 Punkte)*
- [ ] Mode *(Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.2 Punkte)*
- [x] Maximum *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*

Partial results of min() can be aggregated using min() again, partial results of max() can be aggregated using max() again.

For the mean, another aggregate needs to be stored for each partial result, e.g. the size of the subset (if subsets have different sizes). Thus, the mean is algebraic, not distributive.

For computing mode and median, the frequency of each unique value in a subset is needed, meaning that in the worst case, each value needs to be stored anyway.

Frage 6 - Central Tendency (1 Punkt) [ID: 1051778]

Let a feature with the following values be given: 1, 1, 2, 2, 2, 16, 16, 16, 16.

Compute various measures of the central tendency! If necessary, use the decimal point (.) as decimal separator.

The arithmetic mean is 8 *(0.25 Punkte)* .

The median is 2 *(0.25 Punkte)* .

The midrange is 8.5 *(0.25 Punkte)*

The mode is 16 *(0.25 Punkte)* .

This shows how strongly measures of central tendency can differ from each other, depending on the distribution of the data.

The arithmetic mean is what we commonly call "average".

The median is the 50% quantile (50% of the values are lower, 50% are higher).

The midrange is the average of minimum and maximum.

The mode is the most frequent value.

Frage 7 - Boxplot (1 Punkt) [ID: 1051774]

Which quantiles are shown in a boxplot?

- ☑ 0% *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*
- ☑ 25% *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*
- ☑ 50% *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*
- ☑ 75% *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*
- ☑ 100% *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*

By the way, 0% quantile and 100% quantile are minimum and maximum, respectively (in case you didn't know it already ...). Depending on the distribution of the data, min and max might be marked by the position of the whiskers or they might be outlier points. Outliers might also be hidden, depending on the plotting options.

Frage 8 - Entropy (1 Punkt) [ID: 1051780]

Given a feature in which n different values appear with equal frequency, what is the feature's entropy, assuming we use the base-2 logarithm?

- ○ 0
  (0 Punkte)
- ○ 1
  (0 Punkte)
- ○ $\frac{1}{n} * log_2 n$
  (0 Punkte)
- ◉ $log_2 n$
  (1 Punkt)
- ○ $n * log_2 n$
  (0 Punkte)

$H(S) = -\sum_j p_j * \log_2 p_j = -\sum_{i=1}^{n} \frac{1}{n} * \log_2 \frac{1}{n} = -n * (\frac{1}{n} * (\log_2 1 - \log_2 n)) = -1 * (0 - \log$

Frage 9 - Correlation and Independence (1 Punkt) [ID: 1051775]

Which of the following statements is true for two random variables?

- ○ not (Pearson) correlated $\iff$ independent
  (0 Punkte)
- ○ not (Pearson) correlated $\implies$ independent
  (0 Punkte)
- ◉ independent $\implies$ not (Pearson) correlated
  (1 Punkt)
- ○ none of the other statements
  (0 Punkte)

(Pearson) correlation is a special case of a dependency, i.e., if there is no dependency, then there also is no correlation. The other direction does not always hold, as there are complex dependencies that are not captured by simple correlation coefficients like the Pearson correlation coefficient.

Frage 10 - Null Hypotheses for Statistical Tests (1 Punkt) [ID: 1051776]

Match the following null hypotheses to their corresponding statistical tests (according to the definitions from the lecture)!

Random variables are independent.	passt zu	Chi-squared test	*(0.34 Punkte)*
Distributions do not differ.	passt zu	Kolmogorov-Smirnov test	*(0.33 Punkte)*
Medians of the distributions do not differ.	passt zu	Wilcoxon-Mann-Whitney test	*(0.33 Punkte)*

See the lecture. Literature might associate different null hypotheses with the tests.

Frage 11 - Data Reduction (1 Punkt) [ID: 1051779]

Match the types of data reduction to some basic techniques!

Numerosity reduction	passt zu	Sampling	*(0.2 Punkte)*
Numerosity reduction	passt zu	Clustering	*(0.2 Punkte)*
Dimensionality reduction	passt zu	Feature selection	*(0.2 Punkte)*
Dimensionality reduction	passt zu	PCA	*(0.2 Punkte)*
Discretization	passt zu	ChiMerge	*(0.2 Punkte)*

See the lecture.

Though one typically clusters data objects (based on the features), clustering features (based on the data objects) is possible as well, but this is not a common data-reduction technique.

Frage 12 - PCA (1 Punkt) [ID: 1059639]

Which of the following statements on PCA are true?

- ☐ PCA is a supervised technique.
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- ☐ PCA selects a subset of features.
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- ☑ The features after transformation are uncorrelated (in terms of Pearson correlation).
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)
- ☑ The eigenvalues are proportional to the explained variance.
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)

As PCA does not use the target variable (it only depends on the features), it is an unsupervised technique.

PCA does not select features, as the principal components constitute linear combinations of all original features (even though individual features might have more or less influence in each component).

As the covariance matrix is symmetric and only contains real values, the eigenvectors of this matrix (found by PCA) form an orthogonal basis for transformation. This means that the eigenvectors of the covariance matrix have (pairwise) dot products of zero. This implies that the (pairwise) dot products between the transformed features also are zero. As we typically set the mean of the features to zero by normalization

(before conducting PCA), the (pairwise) covariance between transformed features is zero as well (covariance depends on the dot product and the features' means, since Cov(X,Y) = E(X*Y) -E(X)*E(Y)). As a consequence, the (pairwise) Pearson correlation coefficient between transformed features (defined as covariance divided by the individual variances) also becomes zero. In contrast, the eigenvectors of the covariance matrix usually don't have a mean of zero and have (pairwise) correlations coefficients not equal to zero.

The fraction of variance captured in the first **c** principal components equals the sum of the eigenvalues of these principal components divided by the sum of all eigenvalues.

Frage 13 - Normalization (1 Punkt) [ID: 1059637]

After min-max normalization to [0,1], which statistics are always 0.5?

- ☐ Mean
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- ☐ Median
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- ☑ Midrange
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)
- ☐ Mode
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)

Midrange always is in the middle between min and max. All other of the mentioned statistics depend on the distribution of the data points within the interval.

Frage 14 - Search in k-d Trees (1 Punkt) [ID: 1051784]

Which leaf nodes are extracted (i.e., their individual points are added to the search queue) in the nearest-neighbor search in k-d trees?

- ○ Only the leaf node that contains the actual nearest neighbor of the query point.
  (0 Punkte)
- ○ Only the leaf node that contains the query point.
  (0 Punkte)
- ○ The leaf node containing the query point and the leaf node containing the actual nearest neighbor of the query point.
  (0 Punkte)
- ◉ All leaf nodes that are within the distance of the actual nearest neighbor of the query point.
  (1 Punkt)
- ○ All leaf nodes.
  (0 Punkte)

If several leaf nodes have a similar distance to the query point, all of them need to be inspected in detail, as they might contain the nearest neighbor. Therefore, the leaf nodes in the NN sphere, i.e., within the actual NN distance, need to be extracted. Of course, it is not known ex ante how large the NN distance is ...

Frage 15 - Balancing of Spatial Index Structures (1 Punkt) [ID: 1051786]

Which of the following spatial index structures are balanced?

- [x] B tree
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)
- [ ] k-d tree
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- [x] K-D-B tree
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)
- [x] R tree
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)

See the lecture.

Frage 16 - Instance-based Learning (1 Punkt) [ID: 1051785]

Which of the following prediction models are instance-based?

- [ ] Decision tree
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- [ ] Linear classifier
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- [x] k-nearest neighbors
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)
- [ ] One Rule
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)

Decision trees, linear classifiers, and One Rule train a model based on the entire dataset. K-nearest-neighbors-approaches, in contrast, directly search the neighboring data objects without training a model. However, if many of such requests should be made, it makes sense to create a spatial index structure first. This means that instance-based approaches may still have an upfront effort to speed up predictions later.