Online-Test 2: Statistische Grundlagen und Informatik-Grundlagen

Test und Assessment – Druckansicht

Online-Test 2: Statistische Grundlagen und Informatik-Grundlagen

Datum: Tue Nov 2 07:55:17 2021 Maximale Punktezahl: 10

Frage 1 - Kategorisierung von Daten (1 Punkt) [ID: 927162]

Ordnen Sie die Attribute den Datentypen zu!

nominal	passt zu"Farbe", Wertebereich = {grün, blau, rot, gelb, ...}	(0.25 Punkte)
ordinal	passt zu"Note", Wertebereich = {sehr gut, gut, befriedigend, ausreichend, mangelhaft}	(0.25 Punkte)
diskret	passt zu"Würfelzahl", Wertebereich = {1, 2, 3, 4, 5, 6}	(0.25 Punkte)
kontinuierlich	passt zu"Kontostand", Wertebereich = positive reelle Zahlen mit zwei Nachkommastellen	(0.25 Punkte)

"Note" kann natürlich auch auf numerische Art diskret codiert werden.

Frage 2 - Lageparameter (1 Punkt) [ID: 927166]

Gegeben sei ein Attribut mit folgenden Werten: 1, 1, 2, 2, 2, 16, 16, 16, 16

Das arithmetische Mittel beträgt 8 *(0.34 Punkte)* .

Der Median beträgt 2 *(0.33 Punkte)* .

Der Modalwert beträgt 16 *(0.33 Punkte)*

Hier zeigt sich gut, wie stark sich verschiedene Lageparameter, die hier eigentlich alle eine Art Mittelwert im Datensatz berechnen sollen, unterscheiden können.

Frage 3 - Datenreduktion (1 Punkt) [ID: 927171]

Ordnen Sie die Methoden der Datenreduktion den grundlegenden Arten der Datenreduktion zu!

Numerosity Reduction	passt zuSampling	(0.2 Punkte)
Numerosity Reduction	passt zuClustering	(0.2 Punkte)
Dimensionality Reduction	passt zuFeature Selection	(0.2 Punkte)
Dimensionality Reduction	passt zuPCA	(0.2 Punkte)
Diskretisierung	passt zuChiMerge	(0.2 Punkte)

Siehe Vorlesung.

Frage 4 - Entropie (1 Punkt) [ID: 927178]

Gegeben sei ein Datenbestand, in welchem n verschiedene Werte mit jeweils gleicher Häufigkeit auftreten. Was ist die Entropie dieses Datenbestandes?

- ○ 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)

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

Frage 5 - Bernoulli-Experiment (1 Punkt) [ID: 927180]

Welche Verteilung nehmen wir in der Vorlesung an, um den z-Wert für das Konfidenzintervall eines Bernoulli-Experiments zu bestimmen?

- ○ Bernoulli-Verteilung
  (0 Punkte)
- ○ Binomial-Verteilung
  (0 Punkte)
- ◉ Normal-Verteilung
  (1 Punkt)

Jede einzelne Beobachtung ist Bernoulli-verteilt. Die Summe / der Durchschnitt über Bernoulli-verteilte Zufallsvariablen ist Binomial-verteilt. Wir gehen aber von einer großen Anzahl N von Beobachtungen aus und können daher die beobachtete Erfolgsquote f mit einer Standardnormalverteilung annähern (zentraler Grenzwertsatz).

Frage 6 - Normalisierung (1 Punkt) [ID: 927238]

Nach einer Min-Max-Normalisierung beträgt ...

- ○ ... der Median 0.5.
  (0 Punkte)
- ○ ... der Mittelwert 0.5
  (0 Punkte)
- ◉ ... die Midrange 0.5.
  (1 Punkt)
- ○ ... die Standardabweichung 0.5.
  (0 Punkte)

Die Min-Max-Normalisierung setzt das Minimum auf 0 und das Maximum auf 1. Insofern ist die Midrange, definiert als *(max + min) / 2*, danach bei 0.5. Alle anderen genannten statistischen Größen hängen von mehr Daten als nur dem Minimum und dem Maximum ab -- sie werden auch in das Intervall [0, 1] abgebildet, aber über den exakten Wert lässt sich keine pauschale Aussage treffen.

Frage 7 - kd-Baum - Attributwahl (1 Punkt) [ID: 927265]

Nach welchem Prinzip werden die Attribute für Splits im kd-Baum gewählt?

- ○ Zufällig.
  (0 Punkte)
- ○ Entropie-basiert: Jeder Split soll die Entropie bezüglich der Zielvariable minimieren.
  (0 Punkte)
- ◉ Alternierend: Die Attribute werden gemäß einer festen Reihenfolge durchgegangen.
  (1 Punkt)
- ○ Balancierungs-orientiert: Die Attributwahl soll sicherstellen, dass die Blätter die gleiche Höhe haben.
  (0 Punkte)

Entropie ist ein mögliches Kriterium für die Attribut- und Splitwert-Wahl in Entscheidungsbäumen.

kd-Bäume sind nicht zwangsläufig balanciert; dafür gibt es kdB-Bäume.

Frage 8 - kd-Baum - Suche (1 Punkt) [ID: 927267]

Welche Blattknoten müssen bei der NN-Suche in kd-Bäumen betrachtet werden?

- ○ Nur der Blattknoten, der den tatsächlichen nächsten Nachbarn enthält.
  (0 Punkte)
- ○ Nur der Blattknoten, der den Anfragepunkt enthält.
  (0 Punkte)

- (●) Die Blattknoten, welche in der NN-Sphäre des Anfragepunktes liegen.
  (1 Punkt)
- ( ) Alle Blattknoten.
  (0 Punkte)

Der nächste Nachbar kann sich potentiell in mehreren Blattknoten befinden, wenn diese eine ähnliche Distanz zum Anfragepunkt aufweisen. Dadurch müssen die Blattknoten untersucht werden, die sich in der NN-Sphäre befinden. Im Voraus weiß man aber natürlich nicht, wie groß die NN-Distanz ist ...

Frage 9 - Instanz-basiertes Lernen (1 Punkt) [ID: 927269]

Welche der folgenden Vorhersagemodelle sind Instanz-basiert?

- [ ] One Rules
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.33 Punkte)
- [ ] Entscheidungsbaum
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.33 Punkte)
- [x] Nearest neighbors
  (Ausgewählt = 0.34 Punkte, Nicht ausgewählt = 0 Punkte)

Sowohl Entscheidungsbäume als auch One Rules trainieren Regeln, die auf dem ganzen Trainingsdatenbestand basieren. Nearest-neighbors-Verfahren suchen dagegen direkt die Nachbar-Datenobjekte ab. Wenn viele solcher Anfragen gemacht werden sollen, bietet es sich an, eine räumliche Indexstruktur zu erstellen. Somit kann es auch bei Instanz-basierten Verfahren einen Vorabaufwand geben, der mit dem Trainieren eines Modells vergleichbar ist.

Frage 10 - Vergleich von räumlichen Indexstrukturen (1 Punkt) [ID: 927284]

Welche der folgenden räumlichen Indexstrukturen nutzen achsenparallele Rechtecke, um den Raum zu unterteilen?

- [x] kd-Baum
  (Ausgewählt = 0.34 Punkte, Nicht ausgewählt = 0 Punkte)
- [x] kdB-Baum
  (Ausgewählt = 0.33 Punkte, Nicht ausgewählt = 0 Punkte)
- [x] R-Baum
  (Ausgewählt = 0.33 Punkte, Nicht ausgewählt = 0 Punkte)

Alle drei Strukturen nutzen Rechtecke. Unterschiede gibt es zum Beispiel in Fragen der Balancierung oder der Zulässigkeit von Überlappungen.