Aufgabe 1: [DM-Motivation]

a) Beschreiben Sie kurz, was KDD von OLAP unterscheidet.

Während OLAP schwerpunktmäßig dazu eingesetzt wird, die Entwicklung bestimmter Kenngrößen zu überprüfen und im Vorhinein bekannte Hypothesen zu testen, geht es beim KDD darum, im Vorhinein nicht bekannte Zusammenhänge und Regelmäßigkeiten zu entdecken.

b) Warum werden in der Definition von KDD Muster und Zusammenhänge in für den Menschen interpretierbarer Form angestrebt?

KDD bedeutet „Wissensgewinnung in Datenbanken“, d. h. Ziel ist, dass aus den gefundenen Mustern und Zusammenhängen auf dem Wege der Interpretation durch einen sachkundigen Analysten neue Erkenntnisse, neues Wissen über die Daten gewonnen wird. Wissen und damit das Ergebnis des KDD-Prozesses entsteht also ausschließlich durch menschliche Interpretation.

Aufgabe 2: [DM-Aufgaben]

a) Nennen und erklären Sie kurz die drei Eigenschaften, mit denen sich Data-Mining-Aufgaben beschreiben lassen.

Format der Wissensrepräsentation*: Beschreibt, wie das Wissen repräsentiert werden soll, z.B. durch Regeln, Formeln/Modelle, Beispielinstanzen, Kategorien*
Lernart*: überwachtes und unüberwachtes Lernen, siehe Teilaufgabe c)*
Format der Ein- und Ausgabewerte*: Es werden verschiedene Formate der Ein- und Ausgabewerte unterschieden, z.B. numerisch (kontinuierlich, diskret), textuell (nominal, kategorisch, linear)*

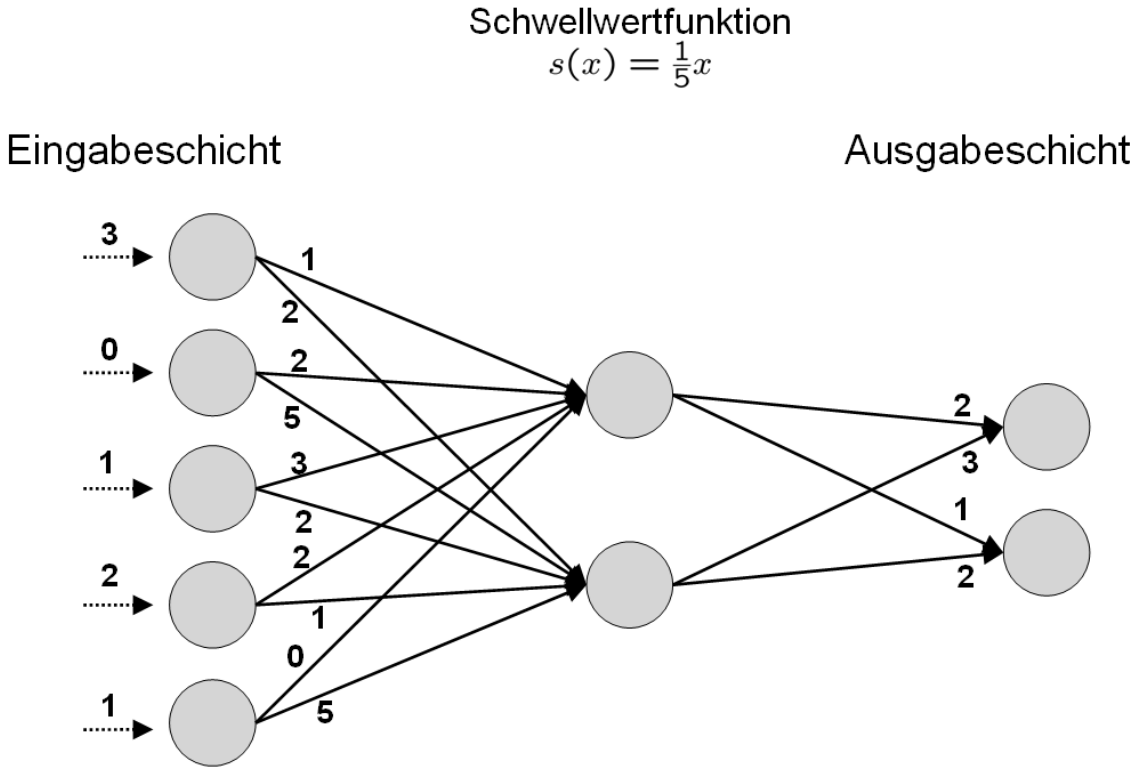
b) Was ist der Unterschied zwischen kategorischen und linearen Werten?

Das Format der Ein- und Ausgabewerte wird als kategorisch bzw. linear bezeichnet, wenn es sich um textuelle, in der Anzahl beschränkte Ausprägungen handelt. Bei einem linearen Format liegt zusätzlich noch eine Ordnung der Ausprägungen vor.
kategorisch*: rot, grün, blau*
linear*: kalt, lauwarm, warm, heiß*

c) Erläutern Sie den Unterschied der in der Vorlesung behandelten Lernarten? Ordnen Sie drei Ihnen bekannte Data-Mining-Aufgaben den verschiedenen Lernarten zu.

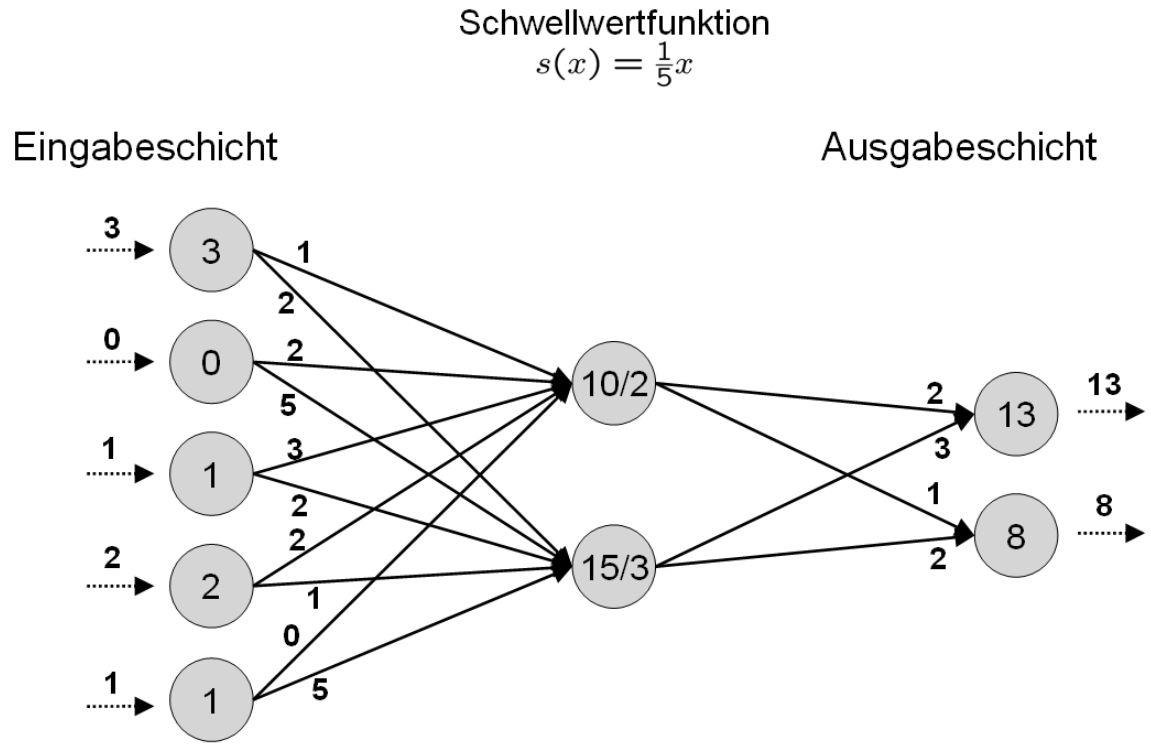
Bei überwachtem Lernen wird das Lernverfahren anhand von Trainingsdaten, die auch entsprechende Ergebnisse enthalten, auf seine Aufgabe vorbereitet (trainiert), beim unüberwachten Lernen wird das Verfahren direkt auf die zu analysierenden Daten angesetzt.
überwachtes Lernen*: Klassifikation, Numerische Prognose*
unüberwachtes Lernen*: Segmentierung, Clustering, Ähnlichkeitsanalyse, Abweichungsanalyse*

Aufgabe 3: [DM-Verfahren: Neuronale Netze]

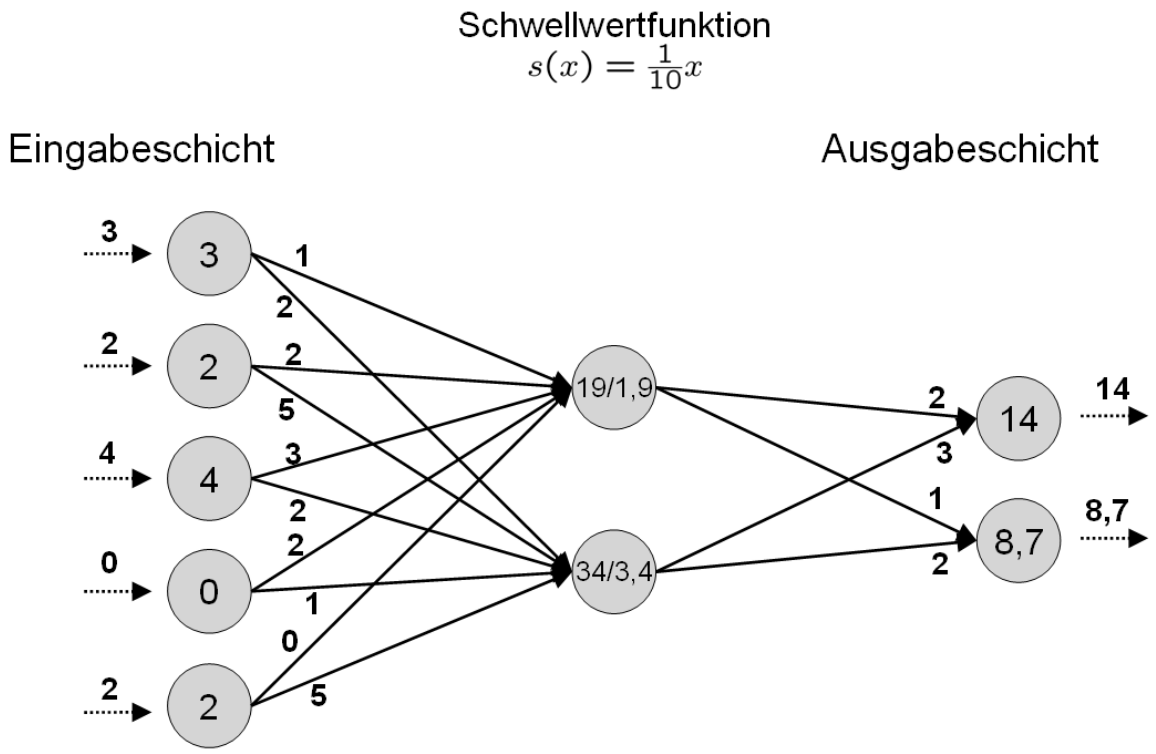


Wenden Sie die Schwellwertfunktion ausschließlich auf die versteckten Schichten an!

a) Vervollständigen Sie das gegebene Neuronale Netz bei Eingabe (3 / 0 / 1 / 2 / 1) und der Schwellwertfunktion $s(x) = \frac{1}{5}x$.



b) Wie muss das Netz aussehen bei Eingabe (3 / 2 / 4 / 0 / 2) und der Schwellwertfunktion $s(x) = \frac{1}{10}x$?



Aufgabe 4: [DM-Verfahren: Assoziationsregeln]

TID	Itemset
1	{1,3,4}
2	{2,3,5}
3	{1,2,3,5}
4	{2,5}

a) Die minimale Unterstützung sei 0,4. Bestimmen Sie für die gegebene Tabelle iterativ die Unterstützung für alle Itemsets und deren Kombinationen.

TID	Itemset
1	{1,3,4}
2	{2,3,5}
3	{1,2,3,5}
4	{2,5}

Itemset	Support
{1}	0,5
{2}	0,75
{3}	0,75
{4}	0,25
{5}	0,75

Itemset	Support
{1}	0,5
{2}	0,75
{3}	0,75
{5}	0,75

Itemset
{1,2}
{1,3}
{1,5}
{2,3}
{2,5}
{3,5}

Itemset	Support
{1,2}	0,25
{1,3}	0,5
{1,5}	0,25
{2,3}	0,5
{2,5}	0,75
{3,5}	0,5

Itemset	Support
{1,3}	0,5
{2,3}	0,5
{2,5}	0,75
{3,5}	0,5

Itemset
{2,3,5}

Itemset	Support
{2,3,5}	0,5

Itemset	Support
{2,3,5}	0,5

b) Wieviele Scan-Vorgänge sind tatsächlich nötig (mit Begründung)?

Es werden drei Scan-Vorgänge benötigt, da nur noch ein drei-elementiges Itemset unterstützt wird und ein vier-elementiges Itemset vier drei-elementige Untermengen besitzt, von denen mindestens zwei nicht unterstützt werden.
oder
Da es nur ein vier-elementiges Itemset und auch nur vier Tupel in der Ausgangsrelation gibt, beträgt die Unterstützung für vier-elementige Itemsets höchstens 0,25 und ist damit kleiner als 0,4.

c) Bestimmen Sie die Konfidenz der Regeln $3 \rightarrow 5$ und $5 \rightarrow 3$.

3 → 5: Konfidenz = $\frac{2}{3}$
5 → 3: Konfidenz = $\frac{2}{3}$

d) Was halten Sie von der Regel $4 \rightarrow 3$ (mit Begründung)?

Diese Regel besitzt zwar hohe Konfidenz, nämlich 1, aber nur geringe Unterstützung, nämlich 0,25. Deswegen ist diese Regel keine geeignete Regel.

Aufgabe 5: [DM-KDD-Prozess]

a) Erläutern Sie, wieso der KDD-Prozess ein iterativer Prozess ist.

Aufgrund der im Vorhinein nicht bekannten Einflussfaktoren auf das letztendliche Ergebnis ist der KDD-Prozess nicht linear und von vornherein planbar, sondern ein explorativer, interaktiver und damit auch iterativer Prozess. In einem iterativen Verfahren werden dabei die untersuchten Datenbereiche, die verwendeten Vorverarbeitungsschritte, die verwendeten Lernverfahren und deren Parameter geeignet eingestellt.

b) Welchen Anteil hat die Phase „Data Understanding“ in einem KDD-Prozess und aus welchen Aufgaben besteht sie?

Data Understanding hat je nach Prozess einen Anteil von 20-30% am KDD-Prozess und dient dazu, sich einen ersten Eindruck über die Daten zu verschaffen. Das bedeutet im Einzelnen, dass die Quellen beschrieben und auf ihre Konsistenz geprüft werden, dass die Bedeutung und Domänen der Attribute analysiert werden (Welche Werte kommen denn so vor?), dass die Daten visualisiert werden und erste offensichtliche Fehler und Mängel in den Daten (fehlende Werte etc.) identifiziert werden.