Übungsblatt LV Künstliche Intelligenz, Data Mining (1), 2014

Aufgabe 1. Data Mining

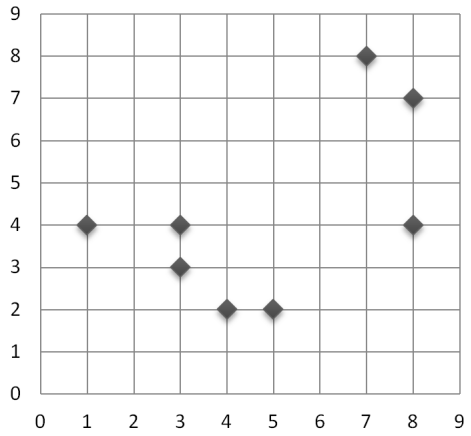
a) Mit welchen Aufgabenstellungen befasst sich Data Mining?

b) Was versteht man unter Transparenz einer Wissensrepräsentation? Nennen Sie je ein transparentes und intransparentes Modell.

Aufgabe 2. Regression, k-means, Assoziationsregeln

a) Ein Modell ist eine vereinfachte Darstellung der Datenmenge. Begründen Sie, inwiefern eine Regressionsgerade "einfacher" ist, als die ursprüngliche Datenmenge.

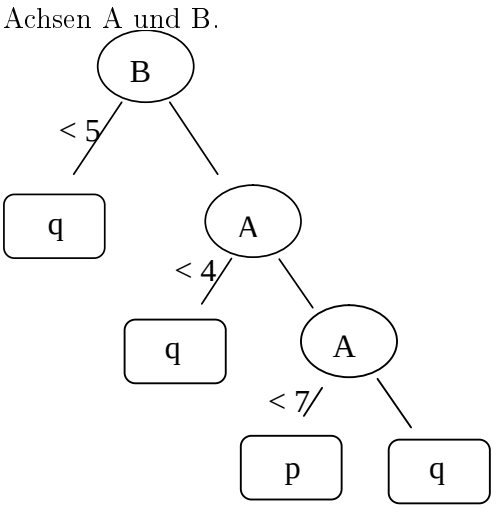
b) Markieren Sie die beiden Cluster, die das k-means-Verfahren mit den initialen Clusterzentren $c_1 = (3, 2)$ und $c_2 = (6, 2)$ findet:



c) Bestimmen Sie Support und Konfidenz der Assoziationsregel $Bier \Longrightarrow Windeln$ in der folgenden Menge beobachteter Warenkörbe: {(Bier,Kakao,Cola), (Cola,Selter,Bier,Windeln), (Kekse,Selter,Bier), (Bier,Kakao,Windeln), (Käse,Bier,Windmühle), (Windeln,Cola,Bier)}.

Aufgabe 3. Zerlegung des Merkmalsraumes durch einen Entscheidungsbaum

a) Skizzieren Sie die Zerlegung des Merkmalsraumes durch folgenden Entscheidungsbaum (A und B sind *Merkmale*, p und q sind *Klassen*). Sie können bspw. ablesen *"Ist B >= 5 und A<4, dann ist die Klasse q"*. Der Merkmalsraum hat die beiden



Aufgabe 4. Entropie

Bei der Konstruktion von Entscheidungsbäumen aus einer Datenmenge beginnen wir mit der Wurzel des Baumes. Hier soll das Attribut stehen, das uns einen hohen Informationsgewinn liefert, wenn sein Attributwert bekannt wird - wir fragen also nach einer Eigenschaft, die uns "viel" verrät. Wir nutzen dafür den Informationsgehalt vor und nach Kenntnis des Attributwertes. Wir betrachten dazu die Häufigkeiten, mit der die einzelnen Attributwerte (bspw. $\{rot, blau, grün\}$) eines Attributes (bspw. *Farbe*) in der Datenmenge auftreten. Diese Häufigkeiten bilden eine (empirische) *diskrete Verteilung*.

Der mittlere Informationsgehalt einer Verteilung S wird als *Entropie* $H(S)$ bezeichnet. Beschreiben wir die Verteilung S durch die relativen Häufigkeiten p_i ihrer Elemente $S = (p_1, ..., p_n)$, so berechnet sich die Entropie von S wie folgt:

$$H(S) = -\sum_{i=1}^{n} p_i \, log_2 \, p_i$$

a) Symbolquelle A erzeugt zufällig einen Zeichenstrom mit dem Alphabet $\{a, b\}$, wobei a dreimal so häufig wie b vorkommt. Symbolquelle B verfügt über das Alphabet $\{x, y, z\}$, wobei $p(x) = p(z) = 0.3$. Welche Symbolquelle hat den höheren Informationsgehalt?

b) Berechnen Sie die Entropie folgender Menge: {Bestseller, Ladenhüter, Bestseller, Ladenhüter, Bestseller, Ladenhüter, Ladenhüter, Ladenhüter}

c) Je nach Geschmack wird ein Produkt zum Ladenhüter oder Bestseller. Bestimmen Sie die Entropien der drei Attributwerte für die folgende Datenmenge.

Geschmack	*Klasse*
salzig	Bestseller
salzig	Ladenhüter
umami	Bestseller
umami	Ladenhüter
umami	Bestseller
umami	Ladenhüter
bitter	Ladenhüter
bitter	Ladenhüter

d) *Freiwillige, prüfungsrelevante Zusatzaufgabe* : Wie groß ist in die *erwartete Entropie* des Attributes Geschmack?

Die erwartete Entropie ist der *Erwartungswert* der Entropie. Der Erwartungswert $E(X)$ einer Zufallsgröße X, die zufällig die Werte $x_1, ...x_n$ annimmt, wird berechnet durch:

$$E(X) = \sum_i p(x_i)\, x_i$$

Hierbei sind $p(x_i)$ die Wahrscheinlichkeiten (hilfsweise relativen Häufigkeiten) der x_i. Wir haben die drei Entropien $H(salzig)$, $H(umami)$ und $H(bitter)$ schon bestimmt und sehen sie als Ausprägung einer Zufallsgröße Entropie, also $x_1 = H(salzig)$, $x_2 = H(umami)$ und $x_3 = H(bitter)$. Mit welcher Häufigkeit tritt x_1 auf, also wie viele Datensätze sind 'salzig' oder wie oft werden wir $H(salzig)$ bei dieser Datenmenge in Anspruch nehmen? Aus den drei Häufigkeiten und den drei Entropien wird so die erwartete Entropie bestimmt.

Übungsblatt LV Künstliche Intelligenz, Data Mining (2), 2014

Fertigen Sie während der Übung ein formloses Protokoll mit den in den Aufgaben genannten Ergebnissen an, Namen im Protokoll nicht vergessen. Laden Sie Ihr fertiges Protokoll heute oder morgen in Moodle hoch.

Aufgabe 1. Datenmenge laden, visualisieren, speichern

Starten Sie Rapidminer (Alle Programme -> RapidMiner 5 -> RapidMiner 5), nehmen Sie keine Updates an. Erzeugen Sie einen *neuen Prozess* (Ctrl-N). Legen Sie dann links im Reiter 'Repositories' mit dem Plus-Icon ein *neues lokales Repository* mit dem Namen 'LV KI 2014' und dem Root Directory C:\Users\pchs\Documents an. Speichern (Disketten-Icon) Sie Ihren leeren Prozess in dem neuen Repository unter dem Prozessnamen 'Mein erster ID3'.

a) Laden der Datenmenge weather.nominal.arff (Operator *Read ARFF*, File aus Moodle zuvor lokal speichern).

b) Anzeige des Streudiagramms (Scatterplot) der Attribute play und outlook. Erkennen Sie den reinen Knoten bei outlook=overcast?
⇒Protokoll 1: Grafik des Streudiagrammes

c) Schreiben Sie die Datenmenge als XLS-File (*WriteExcel*).
⇒Protokoll 2: Tabelle der Daten in Excel oder Calc

Aufgabe 2. Entscheidungsbaumlernen

- Erstellen Sie mit RapidMiner zur Datenmenge weather.nominal.arff einen Entscheidungsbaum mit ID3 (*Read ARFF*, *Set Role*, *ID3*). Den *Set Role*-Operator benötgen Sie, um die Klassenspalte (sog. label) festzulegen.
⇒Protokoll 3: XML des Setups, Grafik des Setups
⇒Protokoll 4: Visualisierung des Modells (=Grafik des Baumes)

Aufgabe 3. Modell anwenden

- Wenden Sie das Modell auf die Daten 2classify.arff an und speichern das Ergebnis als 2classify_classified.xls (*Laden Trainingsmenge*, *Label setzen*, *ID3*, *Laden Testmenge*, *Modell anwenden*, *Excel speichern*). Wie kann die Prognose (prediction) erklärt werden?
⇒Protokoll 5: Grafik des Setups
⇒Protokoll 6: klassifizierte Testmenge (Inhalt xls-File), Prognose bitte markieren
⇒Protokoll 7: Begründung der Prognose mit Hilfe des Baumes

Aufgabe 4. Entscheidungsbaum der Tiere (Daten zoo_german.arff)

1. Erstellen Sie einen Entscheidungsbaum (*Decision Tree*) zur Prognose der Spalte type.
   ⇒Protokoll 8: Grafik des Modells

2. Ist das Modell transparent und plausibel?

   ⇒Protokoll 9: Modell transparent und plausibel?

3. Erstellen Sie einen Baum der maximalen Tiefe 4.
   ⇒Protokoll 10: Grafik des Modells

4. Welcher der beiden Bäume ist *besser*?
   ⇒Protokoll 11: Kurze und präzise Antwort.

Aufgabe 5. *Zusatzaufgabe: Eigenes Data Mining*

1. Welches ist das wichtigste Merkmal einer Region im Satellitenbild einer Stadt, um zu erkennen, ob es sich um einen Pool, ein Gebäude, eine Betonfläche oder was auch immer handelt? Lesen Sie das Attribut aus einem Entscheidungsbaum ab, den Sie aus dem File "training.csv" aus dem Data Set "Urban Land Cover" aus der Datensammlung "UCI Machine Learning Repository" erstellen.
   ⇒Protokoll 12: Kurze und präzise Antwort.

2. Glückwunsch, Sie haben soeben in Aufgabe 3 Ihr erstes Modell aus Daten erstellt und dieses in Aufgabe 4 auf unbekannte Daten angewendet und diese damit vorhergesagt! Welche Datenmenge aus Ihrem persönlichen Umfeld eignet sich vielleicht ebenfalls zur Modellierung, um dann mit dem erstellten Modell (bspw. einem Entscheidungsbaum) etwas bisher Unbekanntes vorherzusagen?

   ⇒Protokoll 13: Meine Idee: Datenmenge, Modell für welche Klasse, Nutzen

Übungsblatt LV Künstliche Intelligenz, Data Mining (3), 2014

Fertigen Sie während der Übung ein formloses Protokoll mit den in den Aufgaben genannten Ergebnissen an, Namen im Protokoll nicht vergessen. Laden Sie Ihr fertiges Protokoll heute oder morgen in Moodle hoch.

Aufgabe 1. Kreuzvalidierung

Erläutern Sie Zweck und Ablauf der stratifizierten Kreuzvalidierung (engl. stratified cross validation), wie heißt der Operator in RapidMiner? Worin besteht der Unterschied zwischen der normalen und der stratifizierten Kreuzvalidierung?
⇒Protokoll 1: Zweck
⇒Protokoll 2: Ablauf
⇒Protokoll 3: Operatorname
⇒Protokoll 4: Unterschied

Aufgabe 2. Resubstitutions-Fehler

Der Resubstitutions-Fehler gibt die Fehlerrate auf den Trainingsdaten an.

1. Erstellen Sie mit RapidMiner den Entscheidungsbaum nach ID3 zur Datenmenge weather.nominal.arff.

2. Bestimmen Sie die Konfusionsmatrix, Fehlerrate (classification_error) und Erfolgsrate (accuracy) des erstellten Modells (dazu im Operator *ClassificationPerformance* den Messwert *accuracy* aktivieren).
   ⇒Protokoll 5: Grafik (=View) des Setups
   ⇒Protokoll 6: Konfusionsmatrix
   ⇒Protokoll 7: Fehlerrate = ..., Erfolgsrate = ...

3. Ist der Resubstitutions-Fehler, den sie gerade bestimmt haben, eine gute Schätzung der Fehlerrate auf unbekannten Daten aus dem Diskursbereich? Schätzt er zu optimistisch oder zu pessimistisch?
   ⇒Protokoll 8: Kurze, präzise Antwort:

Aufgabe 3. Erfolgsrate schätzen

Bestimmen Sie mithilfe der 5-fachen Kreuzvalidierung einen Schätzwert für die Erfolgsrate (*accuracy*) des ID3-Verfahrens mit dem Kriterium *information_gain* auf der Datenmenge weather.nominal.arff.

Hinweis: die Lernvorgang (ID3) und der Messvorgang (*OperatorChain* aus Anwenden und Performance) werden als innere Operatoren der Kreuzvalidierung eingehängt.
⇒Protokoll 9: Grafik (=View) des Setups
⇒Protokoll 10: Erfolgsrate (ID3) = ...

Aufgabe 4. Neuronales Netz

Erreicht ein neuronales Netz (*NeuralNet*) eine bessere Erfolgsrate? Eventuell müssen die nominalen Attribute in numerische Attribute gewandelt werden (*Nominal to Numerical*)

⇒Protokoll 11: Erfolgsrate (KNN) = ...

Aufgabe 5. *Zusatzaufgabe*: Entropie und Performance (alte Prüfungsaufgabe)

1. A und B sind Attribute mit den Attributwerten a1 und a2 bzw. b1 und b2. Es gibt drei Klassen K1, K2 und K3. Bestimmen Sie die *erwartete Entropie des Attributes A* bei folgender Datenmenge:

A	B	Klasse (real)	Klasse (predicted)
a2	b2	K1	
a2	b2	K1	
a2	b1	K3	
a1	b2	K2	
a2	b1	K3	
a2	b1	K3	
a2	b1	K3	
a2	b1	K2	
a2	b2	K2	

   ⇒Protokoll 12: $H(a_1)$ =..... $H(a_2)$ =.....
   ⇒Protokoll 13: $E(A)$ =.....

2. Gegeben sei ein Klassifikator, der die Klasse K3 ausgibt, wenn B den Wert b1 annimmt. Allen anderen Fällen ordnet er die Klasse K2 zu. Wie lautet die Konfusionsmatrix, wenn wir diesen Klassifikator auf die Daten der Tabelle anwenden?
   ⇒Protokoll 14: Konfusionsmatrix

3. Bestimmen Sie die Fehlerrate und Erfolgsrate dieses Klassifikators.
   ⇒Protokoll 15: Fehlerrate = ..., Erfolgsrate = ...