Fragenkatalog zur Vorlesung "Grundlagen des Data Mining" (WS 2006/07)

1. Grenzen Sie die Begriffe "Daten" und "Wissen" mit je 3 charakteristischen Eigenschaften gegeneinander ab.

2. Nennen Sie vier verschiedene Bewertungskriterien für Wissen.

3. Nennen Sie mindestens vier Verarbeitungsschritte in der KDD- (Knowledge Discovery in Databases-) Pipeline.

4. Nennen Sie 3 praktische Anwendungsgebiete des Data Mining und jeweils den potenziellen Nutzen, den das Data Mining speziell auf diesen Gebieten verspricht.

5. Wodurch unterscheiden sich nominalskalierte, ordinalskalierte, intervallskalierte (metrische) und ratioskalierte Attribute?

6. Wann wird ein "missing value" als "missing completely at random" bezeichnet, wann als "missing at random"?

7. Was versteht man unter einem "nonignorable missing value"?

8. Nennen Sie 4 theoriegeleitete und 4 datengeleitete Methoden des Data Mining.

9. (a) Durch welche mathematischen Eigenschaften wird eine Relation "$\leq$" als *Halbordnung* (auch: partielle Ordnung), durch welche als *totale Ordnung* charakterisiert?
(b) Wie lässt sich die Beziehung "a ist unterer Nachbar von b" (bezüglich einer Halbordnung "$\leq$") mathematisch präzise definieren?

10. Wie sind die Begriffe *Kette* und *Antikette* in einer partiell geordneten Menge definiert?

11. Wie sind *Länge* und *Weite* einer endlichen, partiell geordneten Menge M definiert?

12. Was versteht man unter dem *verallgemeinerten Intervall* $[B, C]$ zwischen 2 Mengen B und C innerhalb einer partiell geordneten Menge?

13. Wie sind *Infimum* und *Supremum* einer Menge T innerhalb einer partiell geordneten Menge definiert?

14. Wann heißt eine partiell geordnete Menge ein *Verband*, wann sogar *vollständiger Verband* ?

15. Gegeben sei ein Objektraum $\Omega = X_1 \times ... \times X_n$. Wie ist der Hypothesenraum H über Ω definiert? Wann erfüllt ein Objekt $x \in \Omega$ eine Hypothese $h \in H$?

16. Wie ist die partielle Ordnung auf dem Hypothesenraum H über dem Objektraum $\Omega = X_1 \times ... \times X_n$ definiert (Generalisierungs-Halbordnung) ?

17. Wann ist eine Hypothese *h konsistent* mit einem Lerndatensatz (bestehend aus Positiv- und Negativbeispielen)?

18. Zeichnen Sie ein Liniendiagramm bzgl. der Generalisierungs-Halbordnung für die folgenden Hypothesen, die als partielle Attributbelegungen dargestellt sind:

a = (*unsicher gebunden*, ?, *introvertiert*, *neurotisch*, *männlich*)
b = (*unsicher gebunden*, *sprachgestört*, ?, *neurotisch*, *männlich*)
c = (*sicher gebunden*, ?, *extrovertiert*, ?, *weiblich*)
d = (?, ?, ?, *psychotisch*, ?)
e = (?, ?, ?, ?, ?)
f = (*unsicher gebunden*, ?, ?, *neurotisch*, *männlich*)

Die allgemeinsten Hypothesen sollen sich im Diagramm ganz unten befinden.

19. Beschreiben Sie das Verfahren "FIND-S" für das induktive Lernen aus einem Lerndatensatz mit Positivbeispielen.

20. Wie ist der *Versionenraum* eines Lerndatensatzes (aus Positiv- und Negativbeispielen) bezüglich eines Hypothesenraums *H* definiert?

21. Welche (speicherplatzeffiziente) Darstellung benutzt man im "Candidate-Elimination"-Lernalgorithmus für den Versionenraum der mit den (positiven und negativen) Lernbeispielen konsistenten Hypothesen?

22. Wie ist der *induktive Bias* eines Lernverfahrens formal definiert?

23. Warum ist ein Lernverfahren ohne induktiven Bias praktisch kaum sinnvoll?

24. Skizzieren Sie den Basis-Algorithmus für die *top-down*-Konstruktion eines Entscheidungsbaumes (Eingabe sind Trainingsdaten in Form einer endlichen Tabelle mit *n* als bekannt angenommenen, diskreten Attributen und einem binären Zielattribut, das nur auf der Trainingsmenge gegeben ist und nach dem im Anwendungsfall klassifiziert werden soll).

25. Nach was für einem Kriterium erfolgt sinnvollerweise im Algorithmus aus Frage 24 in jedem Schritt die Auswahl des Splitattributs? (in Worten; keine Formelangaben erforderlich)

26. Konstruieren Sie aus folgendem Datensatz einen Entscheidungsbaum für das Zielattribut "gefährlich" (mit Genauigkeit 100 % auf dem Datensatz).

Substanz	Aggregatzustand	Farbe	Geruch	gefährlich
1	fest	blau	schwach	nein
2	flüssig	gelb	intensiv	ja
3	flüssig	rot	intensiv	nein
4	flüssig	blau	schwach	nein
5	gasförmig	rot	schwach	nein
6	gasförmig	rot	intensiv	ja

27. Was ist der Unterschied zwischen "preference bias" und "restriction bias" ?

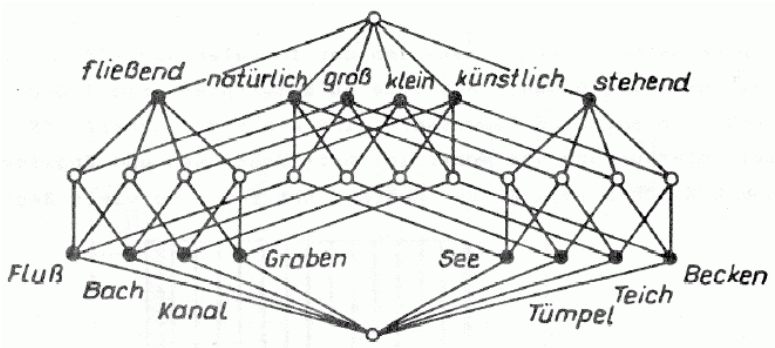
28. Wie würde sich "Overfitting" bei einem Entscheidungsbaum auswirken, und wie kann man es vermeiden?

29. Wie ist der (grobe) Ablauf des "Fehlerreduktions-Prunings" eines Entscheidungsbaumes?

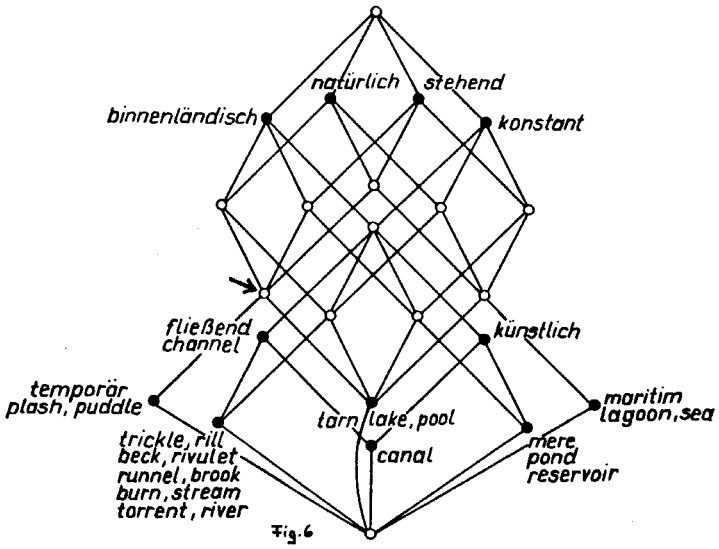
30. Nennen Sie je 4 Vor- und Nachteile des Entscheidungsbaum-Verfahrens beim Data Mining.

31. Wie ist in der formalen Begriffsanalyse ein "Begriff" definiert?

32. Erstellen Sie aus folgendem Liniendiagramm eines Begriffsverbandes eine Kreuztabelle des zugrundeliegenden zweiwertigen Kontexts (Gegenstände: Gewässertypen, Merkmale: Eigenschaften von Gewässern).



33. Welche Gegenstände und welche Merkmale hat der im folgenden begriffsanalytischen Liniendiagramm mit dem Pfeil markierte Begriff?



34. Wie ist der *Gegenstandsbegriff* eines Gegenstandes aus einem Kontext definiert, und welche Minimaleigenschaft hat er?

35. Was versteht man unter "Bereinigung" eines Kontexts?

36. (a) Was versteht man unter einer "Skala" zu einem Merkmal *m* eines mehrwertigen Kontexts?
(b) Wie gelangt man mit Hilfe der Skalenkontexte von einem mehrwertigen Kontext zu einem einfachen Kontext ("schlichte Skalierung")?

37. Skizzieren Sie die Liniendiagramme folgender Elementarskalen: (a) Nominalskala, (b) Ordinalskala, (c) Interordinalskala, (d) Biordinalskala, (e) Boolesche Skala.

38. Skizzieren Sie die grundsätzliche Arbeitsweise des Klassifikationsregel-Lernens mittels des Abdeckungsalgorithmus (sequential covering).

39. Wonach erfolgt im Abdeckungsalgorithmus (zum Klassifikationsregel-Lernen) in jedem Schritt (sinnvollerweise) die Auswahl der neuen Regel (d.h. eines neuen Attributs zum Testen)? (Subroutine "learn-one-rule")

40. Wie unterscheiden sich die Klassifikationsregeln im "Inductive Logic Programming" (ILP) von den "propositional rules" des einfachen Klassifikationsregel-Lernens?

41. Was ist der Unterschied zwischen Klassifikationsregeln und Assoziationsregeln?

42. Was versteht man unter *Support* und *Konfidenz* einer Assoziationsregel?

43. Was liefert der Apriori-Algorithmus als Ergebnis?

44. Was sind die beiden Haupt-Schritte beim Apriori-Algorithmus?

45. Welche Monotonie-Eigenschaft von Item- (Merkmals-) Mengen macht man sich beim Apriori-Algorithmus zunutze?

46. Nennen Sie 2 Nachteile des Apriori-Algorithmus zur Gewinnung von Assoziationsregeln (die Anlass für die Entwicklung verbesserter Verfahren geben).

47. Was versteht man unter einem "anti-monotonen Constraint", und warum lässt sich ein solches bei der Kandidaten-Generierung für Assoziationsregeln verwenden?

48. Was versteht man unter einer *hierarchischen Assoziationsregel*? Geben Sie eine formale Definition.

49. Geben Sie die Aussage des Satzes von Bayes an für die Wahrscheinlichkeit, dass bei vorliegenden Attributwerten $X_1 = x_1, ..., X_n = x_n$ das Zielattribut H den Wert h hat (dass also die Klasse h vorliegt).

50. Gegeben sei ein Hypothesenraum H und ein Datensatz ω. Welche Wahrscheinlichkeit wird
(a) durch die "Maximum a posteriori - Hypothese" aus H für ω,
(b) durch die "Maximum Likelihood - Hypothese" aus H für ω
maximiert?
(c) Unter welcher Voraussetzung sind beide Wahrscheinlichkeiten gleich?

51. Wie lautet die "optimale Bayes-Klassifikationsregel" für ein zu klassifizierendes Objekt x bei Vorliegen des Lerndatensatzes ω ?

52. Welche Annahme wird beim "naiven Bayes-Klassifikator" über die Merkmalswahrscheinlichkeiten gemacht?

53. Was versteht man unter der "Laplace-Schätzung" von Wahrscheinlichkeiten aus Häufigkeiten, und warum wird sie manchmal der einfachen Schätzung (direkt durch die relativen Häufigkeiten) vorgezogen?

54. Nennen Sie je 2 Vor- und Nachteile von Bayes-Klassifikationsverfahren.

55. Was ist die Aufgabe der Clusteranalyse?

56. Was versteht man unter einer *Distanzfunktion* auf einer Menge X ? Geben Sie eine formale Definition.

57. Geben Sie eine Ähnlichkeitsfunktion für Binärstrings an.

58. Welche drei grundlegenden Typen von Clusteranalyse-Verfahren kennen Sie, und welche Parameter sind bei diesen Verfahren jeweils vorzugeben, bevor man die Analyse startet?

59. Nennen Sie 3 verschiedene Varianten von partitionierenden (nichthierarchischen) Clustering-Verfahren. Wodurch wird bei diesen Varianten jeweils ein Cluster repräsentiert?

60. (a) Beschreiben Sie ein Verfahren der partitionierenden, nichthierarchischen Clusteranalyse auf der Grundlage einer Clusterrepräsentation durch Zentroide.
(b) Nennen Sie drei Nachteile dieses Verfahrens.

61. Durch welche Parameter wird eine multivariate Normalverteilung charakterisiert (z.B. im Erwartungsmaximierungs-Algorithmus der Clusteranalyse)?

62. Wie sind in der dichtebasierten Clusteranalyse die "Kernobjekte" definiert?

63. Wie ist ein Dendrogramm zu interpretieren, wenn man daraus eine hierarchische Clusterung ablesen will? Geben Sie ein Beispiel.

64. Beschreiben Sie (in Worten) den Basisalgorithmus der agglomerativen hierarchischen Clusteranalyse. (Die Berechnung von Inter-Cluster-Distanzen muss nicht näher spezifiziert werden.)

65. Was versteht man in der Clusteranalyse unter dem "Chaining-Effekt"?

66. Wie ist die Klassifikationsregel eines "kNN-Klassifikators" (k-nächste-Nachbarn-Klassifikator)?

67. Wie ist die kumulative Verteilungsfunktion eines ordinalskalierten Merkmals definiert?

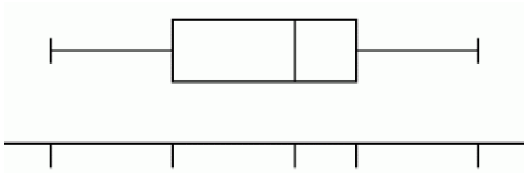
68. Skizzieren Sie jeweils ein Beispiel für die folgenden statistischen Visualisierungstypen:
- Scatterplot,
- bivariates Histogramm,
- Boxplot.

Was ist jeweils auf den Koordinatenachsen aufgetragen?

69. Was versteht man unter der *Kontingenztafel* für 2 Merkmale?

70. Geben Sie die Definitionen der folgenden statistischen Lokalisationsmaße:
- Modalwert
- artithmetischer Mittelwert
- Median.

71. Wie ist in der klassischen Statistik ein *Boxplot* zu interpretieren? Beschriften Sie dazu im folgenden Beispiel die 5 Marker auf der *x*-Achse:



72. Wie sind Varianz und Variationskoeffizient eines reellwertigen Merkmals definiert?

73. Was misst der Korrelationskoeffizient zweier Merkmale, und welchen Wertebereich hat er? Kann man sagen: Wenn der Korrelationskoeffizient zweier Merkmale Null ist, sind diese Merkmale unabhängig?

74. Was versteht man unter dem Signifikanzniveau eines statistischen Tests?

75. Was ist das Ziel der Varianzanalyse?

76. Welches statistische Verfahren bietet sich an für die Klassifikation von Instanzen unter Verwendung einer Lernstichprobe, wenn die Attribute normalverteilte numerische Größen sind?

77. Welche Bedingung sollen die Regressionskoeffizienten beim Verfahren der linearen Regression erfüllen?

78. Was versteht man unter dem *Bestimmtheitsmaß* einer linearen Regression?

79. Warum ist es sinnvoll, im Anschluss an eine lineare Regressionsrechnung eine Analyse (oder zumindest eine Visualisierung) der Residuen durchzuführen?

80. Was ist das Ziel der Faktorenanalyse?

81. (a) Skizzieren Sie ein Beispiel (in Form eines Scatterplots) zur Anwendung der Hauptkomponentenanalyse zur Dimensionsreduktion (von 2 auf 1 Dimension).
(b) Welche Matrix wird zur rechnerischen Durchführung der Hauptkomponentenanalyse benötigt?

82. Was ist das Ziel des *Correlation Clustering* ?

83. Wie werden Texte im "Vektorraummodell" des Text-Minings dargestellt? (einfachster Ansatz, ohne Verwendung von inverser Dokumentenfrequenz oder ähnlicher Verfeinerungen)

84. (a) Was versteht man unter einem "linearen Klassifikator"?
(b) Wie wird beim Verfahren von Rocchio die trennende Hyperebene des linearen Klassifikators zwischen 2 Klassen (von Lerndaten) bestimmt?
(c) Welche Forderung stellt man bei einem "large margin classificator" an die trennende Hyperebene?