Übung 4 Entscheidungsbäume II & Hauptkomponentenanalyse I

Aufgabe 1

Die folgende Tabelle gibt für 4 Kreditantragsteller an, ob sie den beantragten Kredit erhalten haben oder nicht und welche von 3 Risikomerkmalen sie aufweisen.

Antragsteller	RM 1 (X_1)	RM 2 (X_2)	RM 3 (X_3)	Kredit erhalten (T)
1	Ja	Ja	Ja	Ja
2	Ja	Ja	Nein	Nein
3	Ja	Nein	Nein	Ja
4	Nein	Nein	Ja	Nein

a) Konstruieren Sie einen Klassifikationsbaum. Verwenden Sie bei der Bestimmung der Partition $\mathcal{P}$ die Fehlklassifikationsrate $\varepsilon_1(R)$ und als Verlustfunktion die Fehlerfunktion

$$\frac{|R_0(i)|}{|R|} \cdot \varepsilon_1(R_0(i)) + \frac{|R_1(i)|}{|R|} \cdot \varepsilon_1(R_1(i)).$$

Zeichnen und beschriften Sie den resultierenden Klassifikationsbaum.

b) Welche Prognose liefert der in Aufgabenteil a) resultierende Klassifikationsbaum für einen Kreditantragsteller, der das 1. und 3. Risikomerkmal, aber nicht das 2. Risikomerkmal aufweist?

Aufgabe 2

Die folgende Tabelle gibt für 14 Patienten an, ob sie mit einer bestimmten Krankheit infiziert sind und welche von 3 Krankheitssymptomen sie aufweisen.

Patient	Fieber (X_1)	Husten (X_2)	Kurzatmigkeit (X_3)	infiziert (T)
1	Nein	Nein	Nein	Nein
2	Ja	Ja	Ja	Ja
3	Ja	Ja	Nein	Nein
4	Ja	Nein	Ja	Ja
5	Ja	Ja	Ja	Ja
6	Nein	Ja	Nein	Nein
7	Ja	Nein	Ja	Ja
8	Ja	Nein	Ja	Ja
9	Nein	Ja	Ja	Ja
10	Ja	Ja	Nein	Ja
11	Nein	Ja	Nein	Nein
12	Nein	Ja	Ja	Nein
13	Nein	Ja	Ja	Nein
14	Ja	Ja	Nein	Nein

a) Konstruieren Sie einen Klassifikationsbaum. Verwenden Sie bei der Bestimmung der Partition $\mathcal{P}$ die Entropie $\varepsilon_3(R)$ und als Verlustfunktion die Fehlerfunktion

$$\frac{|R_0(i)|}{|R|} \cdot \varepsilon_3(R_0(i)) + \frac{|R_1(i)|}{|R|} \cdot \varepsilon_3(R_1(i)).$$

Zeichnen und beschriften Sie den resultierenden Klassifikationsbaum.

b) Geben Sie für den in Aufgabenteil a) resultierenden Klassifikationsbaum die Fehlklassifikationsrate an.

c) Reduzieren Sie die Varianz des in Aufgabenteil a) erhaltenen Klassifikationsbaums durch Reduced error pruning. Verwenden Sie dabei die Fehl- klassifikationsrate als Pruning-Kriterium.

d) Welche Prognose liefert der in Aufgabenteil c) resultierende Klassifikationsbaum für einen Patienten ohne Fieber und ohne Husten, aber mit Kurzatmigkeit?

Aufgabe 3

Die folgende Tabelle gibt für 5 ausgewählte Länder die im Rahmen einer PISA-Studie jeweils erreichte Punktzahl für die beiden ausgewählten Merk- male „Lesekompetenz X_1“ und “Mathematische Grundbildung X_2“ an:

Land	Lesekompetenz (X_1)	Mathematische Grundbildung (X_2)
Dänemark (DK)	497	514
Griechenland (GR)	474	447
Italien (I)	487	457
Portugal (P)	470	454
Schweden (S)	516	510

a) Stellen Sie die Daten in einem Streudiagramm dar und interpretieren Sie das Diagramm.

b) Bestimmen Sie die zentrierte Datenmatrix $\mathbf{D}$, um die Interpretation zu vereinfachen und berechnen Sie die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$.

c) Die beiden Merkmale sollen nun durch Linearkombinationen dieser beiden Merkmale ersetzt werden, die möglichst viel von der in Aufgabenteil a) beobachteten zweidimensionalen Struktur (Varianz) erhalten. Berechnen Sie hierzu die beiden normierten Eigenvektoren von $\mathbf{S}$ und geben Sie die beiden Hauptkomponenten an. Interpretieren Sie ferner das Ergebnis.

Aufgabe 4

Der dreidimensionale Zufallsvektor $\mathbf{X} = (X_1, X_2, X_3)^T$ besitzt die Varianz-Kovarianz-Matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

a) Berechnen Sie die Eigenwerte und zugehörigen normierten Eigenvektoren von $\boldsymbol{\Sigma}$.

b) Geben Sie die Hauptkomponenten des Zufallsvektors $\mathbf{X} = (X_1, X_2, X_3)^T$ an.