

30. Juni 2006	Prüfung aus Einführung in die Mustererkennung — SS 2006	9.30 – 11.00
Matr. Nummer:	Nachname:	
Kennzahl:	Vorname:	

Bei der vorliegenden Prüfung können Sie eine maximale Anzahl von 30 Punkten erreichen. Bitte verwenden Sie den für die Beantwortung der Frage vorgesehenen Platz und beantworten Sie die folgenden Fragen kurz aber aussagekräftig. Sie können die Fragen auf Englisch oder Deutsch beantworten. **Keine Unterlagen sind erlaubt.**

1 Kurze Fragen

1. Ist der k -NN Klassifikator ein Überwachtes oder unüberwachtes Lernverfahren? Die nicht-passende Antwort durchstreichen. (1 Punkt)

Überwachtes Verfahren / Unüberwachtes Verfahren

Begründung: weil das gesamte Trainingsset der Merkmale bereits gegeben ist und nur der Klassenwert fehlt.

2. Ist der k -NN Klassifikator ein parametrisches oder nicht-parametrisches Verfahren? Die nicht-passende Antwort durchstreichen. (1 Punkt)

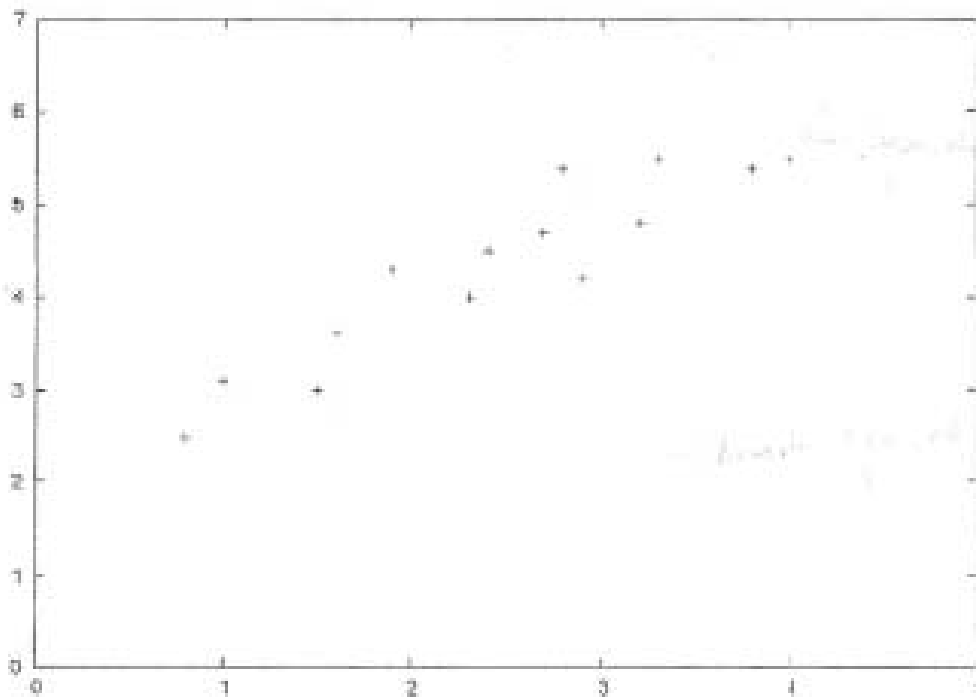
Parametrisches Verfahren / Nicht-parametrisches Verfahren

Begründung: weil keine Annahme über die Form der Entscheidungsgrenze gemacht wird.

3. Unter welchen Bedingungen konvergiert der *online Perceptron Training Algorithmus* (mit fixer Lernrate)?

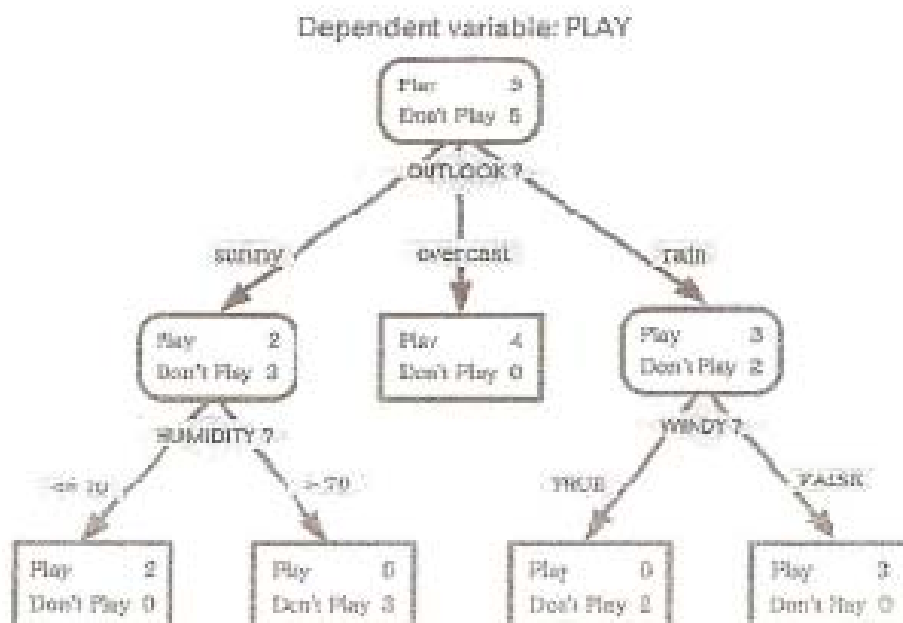
Was macht der Algorithmus, wenn diese Bedingungen nicht erfüllt sind? (1 Punkt)

4. Zeichnen Sie die Orientierungen des Haupteigenvektors und Zweiten Eigenvektors von einer Hauptkomponentenanalyse (Principal Component Analysis, PCA) für den folgenden Datensatz. Die eingezeichneten Richtungen sollen klar beschriftet sein. (1 Punkt)



5. Dieser Entscheidungsbaum entscheidet, ob man Tennis spielen gehen soll oder nicht. Die 4 Merkmale sind (Outlook, Temperature, Humidity, Windy). Welche Entscheidung wird für diese Merkmalsausprägungen getroffen: (sunny, 25, 60, FALSE)? Die Werte in den Knoten geben die Anzahl der Trainingsbeispiele an, die in den jeweiligen Knoten ankommen. Die nicht-passende Antwort durchstreichen. (1 Punkt)

Play / Don't Play



6. Die Aktivierungsfunktion in einem Backpropagation neuronalen Netz muss

differenzierbar / eine Sigmoid Funktion

sein. Die nicht-passende Antwort durchstreichen. (1 Punkt)

7. Die zwei Zufallsvariablen A und B sind statistisch *unabhängig* voneinander. Welche von den folgenden Ausdrücken ist **FALSCH**? Schreiben Sie die richtige Antwort ins untere Rechteck. (1 Punkt)

(a) $P(A|B) = P(A)$

(b) $P(A|B) = P(B)$

(c) $P(A, B) = P(A|B)P(B)$

(d) $P(A, B) = P(B|A)P(A)$

Dieser Ausdruck ist FALSCH:

2 Rechenbeispiele

1. Ein Perceptron (mit einer Sigmoid Funktion) wird durch den folgenden Gewichtsvektor beschrieben:

$$\mathbf{w} = \begin{pmatrix} 0.3 \\ 0.8 \\ 1.7 \\ -0.8 \end{pmatrix} \quad (1)$$

Der folgende Merkmalsvektor ist mit diesem Perceptron zu klassifizieren:

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0.3 \\ -3.8 \\ 0.2 \end{pmatrix} \quad (2)$$

(a) Werden hier homogene or nicht-homogene Koordinaten verwendet? Die nicht-passende Antwort durchstreichen. (1 Punkt)

Homogene Koordinaten / Nicht-homogene Koordinaten

Begründung:

- (b) Was ist der Ausgabewert von diesem Perceptron mit dem gegebenen Merkmalsvektor? Schreiben Sie diesen Wert ins Rechteck unten. (1 Punkt)

$$y = \text{sign}(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6 + w_7 x_7 + w_8 x_8 + w_9 x_9 + w_{10} x_{10} + w_{11} x_{11} + w_{12} x_{12} + w_{13} x_{13} + w_{14} x_{14} + w_{15} x_{15} + w_{16} x_{16} + w_{17} x_{17} + w_{18} x_{18} + w_{19} x_{19} + w_{20} x_{20} + w_{21} x_{21} + w_{22} x_{22} + w_{23} x_{23} + w_{24} x_{24} + w_{25} x_{25} + w_{26} x_{26} + w_{27} x_{27} + w_{28} x_{28} + w_{29} x_{29} + w_{30} x_{30} + w_{31} x_{31} + w_{32} x_{32} + w_{33} x_{33} + w_{34} x_{34} + w_{35} x_{35} + w_{36} x_{36} + w_{37} x_{37} + w_{38} x_{38} + w_{39} x_{39} + w_{40} x_{40} + w_{41} x_{41} + w_{42} x_{42} + w_{43} x_{43} + w_{44} x_{44} + w_{45} x_{45} + w_{46} x_{46} + w_{47} x_{47} + w_{48} x_{48} + w_{49} x_{49} + w_{50} x_{50} + w_{51} x_{51} + w_{52} x_{52} + w_{53} x_{53} + w_{54} x_{54} + w_{55} x_{55} + w_{56} x_{56} + w_{57} x_{57} + w_{58} x_{58} + w_{59} x_{59} + w_{60} x_{60} + w_{61} x_{61} + w_{62} x_{62} + w_{63} x_{63} + w_{64} x_{64} + w_{65} x_{65} + w_{66} x_{66} + w_{67} x_{67} + w_{68} x_{68} + w_{69} x_{69} + w_{70} x_{70} + w_{71} x_{71} + w_{72} x_{72} + w_{73} x_{73} + w_{74} x_{74} + w_{75} x_{75} + w_{76} x_{76} + w_{77} x_{77} + w_{78} x_{78} + w_{79} x_{79} + w_{80} x_{80} + w_{81} x_{81} + w_{82} x_{82} + w_{83} x_{83} + w_{84} x_{84} + w_{85} x_{85} + w_{86} x_{86} + w_{87} x_{87} + w_{88} x_{88} + w_{89} x_{89} + w_{90} x_{90} + w_{91} x_{91} + w_{92} x_{92} + w_{93} x_{93} + w_{94} x_{94} + w_{95} x_{95} + w_{96} x_{96} + w_{97} x_{97} + w_{98} x_{98} + w_{99} x_{99} + w_{100} x_{100})$$

Ausgabewert vom Perceptron:

2. Sie haben zwei Klassen und wollen die Entscheidungsgrenze zwischen diesen zwei Klassen berechnen. Das Trainingsset für jede Klasse besteht aus 4 zwei-dimensionalen Vektoren. Die Trainingsvektoren für Klasse 1 sind:

$$\mathbf{x}_{11} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \mathbf{x}_{12} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \mathbf{x}_{13} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{x}_{14} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \quad (3)$$

und die Trainingsvektoren für Klasse 2 sind:

$$\mathbf{x}_{21} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \mathbf{x}_{22} = \begin{pmatrix} 7 \\ 3 \end{pmatrix}, \mathbf{x}_{23} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \mathbf{x}_{24} = \begin{pmatrix} 6 \\ 4 \end{pmatrix} \quad (4)$$

- (a) Berechnen Sie die geschätzten Mittelwertvektoren $\hat{\mu}_j$ für diese zwei Klassen. Schreiben Sie diese Vektoren in die unteren Rechtecke. (1 Punkt)

$$\mu_1 = \frac{1}{4} \left(\begin{pmatrix} 1 \\ 3 \end{pmatrix} + \begin{pmatrix} 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 2 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right) = \frac{1}{4} \begin{pmatrix} 8 \\ 12 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

$$\mu_2 = \frac{1}{4} \left(\begin{pmatrix} 5 \\ 3 \end{pmatrix} + \begin{pmatrix} 7 \\ 3 \end{pmatrix} + \begin{pmatrix} 6 \\ 2 \end{pmatrix} + \begin{pmatrix} 6 \\ 4 \end{pmatrix} \right) = \frac{1}{4} \begin{pmatrix} 24 \\ 12 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$$

Mittelwertvektor $\hat{\mu}_1 =$

Mittelwertvektor $\hat{\mu}_2 =$

- (b) Berechnen Sie die geschätzte Kovarianzmatrix $\hat{\Sigma}_1$ für Klasse 1. Schreiben Sie diese Matrix ins Rechteck unten. (1 Punkt)

$$\hat{\Sigma}_1 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 \\ x_{i2} - \bar{x}_2 & x_{i3} - \bar{x}_3 \end{pmatrix} \begin{pmatrix} x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 \\ x_{i2} - \bar{x}_2 & x_{i3} - \bar{x}_3 \end{pmatrix}^T$$

$$= \frac{1}{4} \left(\begin{pmatrix} 2-2 \\ 2-3 \end{pmatrix} \begin{pmatrix} 2-2 & 2-3 \end{pmatrix} + \begin{pmatrix} 2-2 \\ 2-3 \end{pmatrix} \begin{pmatrix} 2-2 & 2-3 \end{pmatrix} + \begin{pmatrix} 2-2 \\ 2-3 \end{pmatrix} \begin{pmatrix} 2-2 & 2-3 \end{pmatrix} + \begin{pmatrix} 2-2 \\ 2-3 \end{pmatrix} \begin{pmatrix} 2-2 & 2-3 \end{pmatrix} \right)$$

$$= \frac{1}{4} \left(\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \right) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

Kovarianzmatrix $\hat{\Sigma}_1 =$

$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

- (c) Sie wollen Entscheidungsfunktionen, die auf dem Bayes Theorem basieren, verwenden. Wie berechnen sich gemäß Bayes-Theorem die *a posteriori* Wahrscheinlichkeiten $p(\omega_j | x)$ (1 Punkt)

$$p(\omega_j | x) = \frac{p(x | \omega_j) p(\omega_j)}{p(x)}$$

$$p(\omega_j | x) = \frac{\text{Likelihood} \cdot \text{prior}}{\text{evidenz}}$$

(d) Sie modellieren die Klassen durch zwei-dimensionale Gauß'sche Verteilungen:

$$p(x|\omega_j) = \frac{1}{(2\pi)^{\frac{a}{2}} |\hat{\Sigma}_j|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (x - \hat{\mu}_j)^T \hat{\Sigma}_j^{-1} (x - \hat{\mu}_j) \right] \quad (5)$$

wobei $a = 2$. Schreiben Sie die Diskriminanten-Funktion $g_j(x) = \ln p(\omega_j|x)$ hier aus (die Gauß'sche Verteilung für $p(x|\omega_j)$ soll auch in diesen Ausdruck hineingenommen werden). N.B. Der Ausdruck soll eine Funktion von $\hat{\mu}_j, \hat{\Sigma}_j, \omega_j, P(\omega_j), p(x)$ und x sein. Klassen-spezifischen Werten sollen nicht verwendet werden. (1 Punkt)

$$g_j(x) = -\frac{a}{2} \ln |\hat{\Sigma}_j| - \frac{1}{2} (x - \hat{\mu}_j)^T \hat{\Sigma}_j^{-1} (x - \hat{\mu}_j) + \ln P(\omega_j)$$

$$d_j(x) = (x - \hat{\mu}_j)^T \hat{\Sigma}_j^{-1} (x - \hat{\mu}_j) + a \ln |\hat{\Sigma}_j| - 2 \ln P(\omega_j)$$

(e) Warum kann man die Terme: $-\ln 2\pi - \ln p(x)$ auslassen, ohne dass die Entscheidungsregionen und Entscheidungsgrenzen sich ändern. (1 Punkt)

Da diese Terme konstant sind, ändern sie die Entscheidungsgrenzen nicht. Die Entscheidungsgrenzen werden durch die Diskriminantenfunktion $g_j(x)$ bestimmt, die diese Terme nicht enthält.

- (f) Sie nehmen an, dass $P(\omega_1) = P(\omega_2) = \frac{1}{2}$. Die Diskriminanten-Funktionen für die zwei Klassen schauen so aus:

$$g_1(\mathbf{x}) = -\frac{3}{4}x_1^2 + 3x_1 - \frac{3}{4}x_2^2 + \frac{9}{2}x_2 - \frac{39}{4} - \frac{1}{2}\ln\left(\frac{4}{9}\right) + \ln\left(\frac{1}{2}\right) \quad (6)$$

$$g_2(\mathbf{x}) = -\frac{3}{4}x_1^2 + 9x_1 - \frac{3}{4}x_2^2 + \frac{9}{2}x_2 - \frac{135}{4} - \frac{1}{2}\ln\left(\frac{4}{9}\right) + \ln\left(\frac{1}{2}\right) \quad (7)$$

wobei x_1 und x_2 die zwei Komponenten vom einem beliebigen Vektor \mathbf{x} sind. Berechnen Sie den Ausdruck für die Entscheidungsgrenze. (1 Punkt)

Handwritten work:

$$g_1(\mathbf{x}) - g_2(\mathbf{x}) = \left(-\frac{3}{4}x_1^2 + 3x_1 - \frac{3}{4}x_2^2 + \frac{9}{2}x_2 - \frac{39}{4} - \frac{1}{2}\ln\left(\frac{4}{9}\right) + \ln\left(\frac{1}{2}\right)\right) - \left(-\frac{3}{4}x_1^2 + 9x_1 - \frac{3}{4}x_2^2 + \frac{9}{2}x_2 - \frac{135}{4} - \frac{1}{2}\ln\left(\frac{4}{9}\right) + \ln\left(\frac{1}{2}\right)\right)$$

$$= 3x_1 - 9x_1 = -6x_1$$

- (g) Begründen Sie, warum die von Ihnen berechnete Entscheidungsgrenze richtig ist. (1 Punkt)

- (h) Schreiben Sie den Ausdruck für die Mahalanobis-Distanz d zwischen einem beliebigen Vektor x und der Klasse 1 (mit Mittelwertvektor $\hat{\mu}_1$ und Kovarianzmatrix $\hat{\Sigma}_1$). (1 Punkt)

Handwritten note: $d(x, \hat{\mu}_1, \hat{\Sigma}_1) = \sqrt{(x - \hat{\mu}_1)^T \hat{\Sigma}_1^{-1} (x - \hat{\mu}_1)}$

- (i) Berechnen Sie die Mahalanobis Distanz zwischen Klasse 1 und dem Vektor

$$x_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (8)$$

Schreiben Sie die berechnete Distanz ins untere Rechteck. (1 Punkt)

TIP: Die Inverse von einer 2×2 Matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad (9)$$

ist

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad (10)$$

Handwritten calculation:
 $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$
 $\det(A) = 1 \cdot 1 - 1 \cdot 1 = 0$
 Die Matrix ist nicht invertierbar.

Handwritten calculation:
 $A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{1 \cdot 1 - 1 \cdot 1} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$
 Division durch Null ist nicht möglich.

Handwritten note: $d = \sqrt{(x - \hat{\mu}_1)^T \hat{\Sigma}_1^{-1} (x - \hat{\mu}_1)}$

Mahalanobis Distanz $d =$

3 Detaillierte Fragen

1. Beschreiben Sie kurz das k-NN Verfahren. Welche Vor- bzw. Nachteile besitzt es gegenüber Verfahren wie dem Perceptron oder dem Mahalanobis Distanz Klassifikator? (2 Punkte)
Das k-NN Verfahren ist ein Klassifikationsverfahren, bei dem die Klasse eines neuen Datenpunktes basierend auf der Klasse seiner k nächsten Nachbarn bestimmt wird. Vorteile: Einfachheit, keine Parameter, keine Annahmen über die Datenverteilung. Nachteile: Sensibilität gegenüber Rauschen, hoher Rechenaufwand bei großen Datensätzen.
2. Beschreiben Sie das *Single Linkage Hierarchical Clustering* Verfahren. Wie funktioniert der Algorithmus? Erklären Sie mittels einem 2-dimensionalen Beispiel. (2 Punkte)
Das Single Linkage Hierarchical Clustering Verfahren ist ein hierarchisches Clustering-Verfahren, bei dem die Distanz zwischen zwei Clustern als die Distanz zwischen den beiden nächstgelegenen Punkten der Clustern definiert wird. Der Algorithmus beginnt mit jedem Datenpunkt als eigenem Cluster und verbindet die Clustern schrittweise, bis alle Datenpunkte in einem einzigen Cluster sind. Ein 2-dimensionales Beispiel: Gegeben seien zwei Klassen von Datenpunkten in einem 2D-Raum. Der Algorithmus verbindet zuerst die beiden nächstgelegenen Punkte der beiden Klassen zu einem Cluster und wiederholt dies, bis alle Punkte in einem Cluster sind.
3. Sie haben die Klassenwahrscheinlichkeiten $P(\omega_1|x)$ und $P(\omega_2|x)$ für ein Problem mit zwei Klassen berechnet, wobei x der Merkmalsvektor ist. Für welche Klasse soll, laut *Bayes Decision Rule* (Bayes Entscheidungsregel), entschieden werden? Beweisen Sie, dass die Bayes-Entscheidungsregel die Fehlerwahrscheinlichkeit $P(error)$ im Fall von 2 Klassen immer minimiert. (2 Punkte)
Die Bayes-Entscheidungsregel wählt die Klasse ω_1 aus, wenn $P(\omega_1|x) > P(\omega_2|x)$, ansonsten ω_2 . Die Fehlerwahrscheinlichkeit $P(error)$ ist die Wahrscheinlichkeit, dass die Entscheidung nicht der tatsächlichen Klasse entspricht. Die Bayes-Entscheidungsregel minimiert $P(error)$, da sie die Entscheidung basierend auf den maximalen Posterior-Wahrscheinlichkeiten trifft.

4 Aufsatz

Wählen Sie eines der folgenden Themen und schreiben Sie 100–200 Wörter darüber (6 Punkte). Die Themen sind:

1. PCA in der Gesichtserkennung.
2. Der Aufbau von Entscheidungsbäumen.
3. Bias, Varianz und Generalisierungsfähigkeit.