

DKFZ Heidelberg

Maschinelles Lernen, Klassifikation und diagnostische Tests

Falk Schubert
Intelligente Bioinformatiksysteme
DKFZ Heidelberg

f.schubert@dkfz.de

DKFZ Heidelberg

Prolog

Was ist das?



DKFZ Heidelberg

3

Falk Schubert

Signifikanz?



DKFZ Heidelberg

4

Falk Schubert

Warum?



- Typische Eigenschaften:
 - Farbe, Form, Oberfläche, Größe

Farbe?



Form?



DKFZ Heidelberg

7



Falk Schubert

Oberfläche?



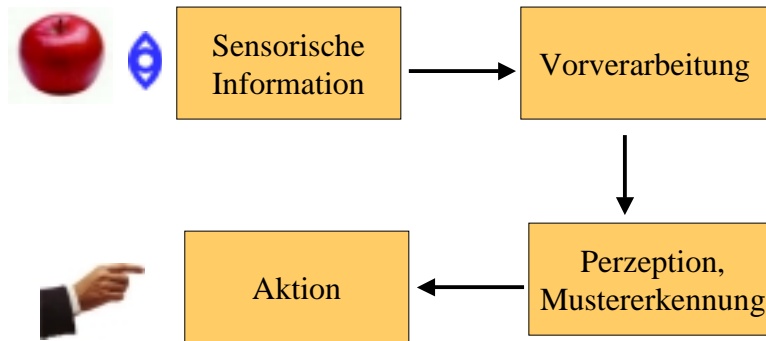
DKFZ Heidelberg

8



Falk Schubert

Menschliche Musterklassifikation

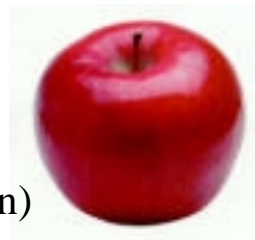


Zusammenfassung

■ Menschlicher Klassifikator:

- Fehlertolerant
- Flexibel
- Klassifikation ist erklärbar

- ## ■ Manchmal ist der menschliche Klassifikator überlegen (Bilderklassifikation)



Erkennung von Gefühlszuständen



Lachen



Angst

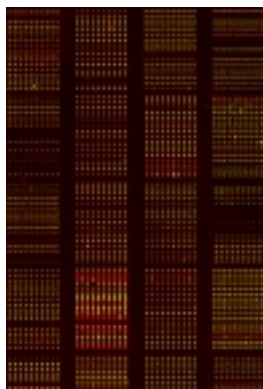
Molekulare Biotechnologie

- Proteine
- Gene
- Genomische Profile
- Genexpressionsprofile

- -> Gene prediction

Klassifikation und molekulare Biotechnologie

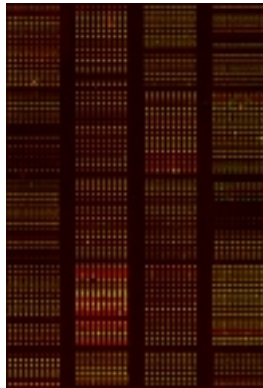
Von welchem Tumortyp ist diese Probe?



0.43	0.38	0.38	0.08	0.25	0.11	0.04	0.33
0.25	0.46	-0.13	0.05	0.15	-0.19	missing	0.46
-0.23	0.18	-0.11	-0.19	0.18	0.28	-0.26	-0.26
0.35	0.32	0.19	0.16	0.03	0.02	-0.08	0.23
0.43	missing	0.26	0.04	0.14	0.11	0.02	0.08
0.48	missing	-0.14	0.23	0.28	0.08	0.08	0.37
0.25	-0.03	-0.17	0.15	-0.17	0.13	-0.14	0.38
0.53	0.44	0.22	0.08	0.46	0.08	0.11	0.09
0.88	-0.12	0.24	0.04	-0.04	0.14	-0.13	-0.27
0.86	0.58	-0.12	-0.03	-0.07	0.12	0.08	-0.03
0.28	0.56	-0.22	0.04	0.07	-0.04	-0.03	0.07
0.89	0.25	0.25	0.03	-0.22	0.07	0.09	0.13
0.11	0.48	0.38	0.02	-0.07	0.26	0.02	0.02
0.17	0.35	0.36	-0.03	-0.06	0.23	0.08	-0.33
0.14	0.48	0.33	-0.03	0.03	0.23	0.11	-0.12
-0.13	0.43	0.32	0.08	-0.26	0.23	-0.17	-0.02
-0.87	0.19	-0.18	0.02	-0.04	0.04	0.08	-0.19
-0.84	0.48	-0.32	-0.03	0.03	-0.03	-0.12	-0.03
-0.35	missing	0.12	0.08	0.36	0.17	-0.14	0.27
-0.38	0.48	-0.19	-0.02	-0.02	0.04	-0.27	0.38
0.84	0.81	-0.37	-0.08	0.26	0.03	-0.27	-0.17
0.88	0.58	0.38	0.11	0.46	0.08	missing	-0.19
0.86	0.24	0.22	0.12	0.05	0.08	-0.22	0.23
-0.84	missing	0.34	0.28	0.22	-0.03	-0.15	-0.07
-0.83	0.22	0.36	0.28	0.28	-0.07	-0.14	0.44
0.83	0.48	-0.28	0.06	0.27	0.02	-0.24	0.08
-0.35	0.38	-0.11	0.09	0.36	-0.08	missing	0.03
0.87	0.43	-0.18	0.03	0.18	0.03	-0.14	missing
-0.25	-0.28	0.34	0.11	0.27	0.03	-0.11	-0.25

matrix-CGH

Welche Eigenschaften kennzeichnen einen Tumortyp?

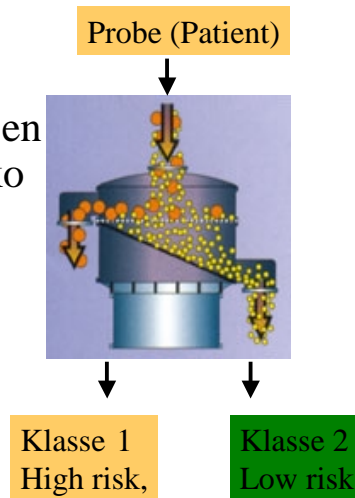


0.43	0.38	0.38	0.38	0.25	0.11	0.04	0.34
0.25	0.46	-0.12	0.08	0.18	-0.19	matrix	0.44
-0.23	0.18	-0.11	-0.18	0.18	0.18	0.28	-0.36
0.35	0.30	0.19	0.18	0.03	0.05	0.08	0.20
0.43	matrix	0.36	0.04	0.14	0.11	0.00	0.08
0.48	matrix	0.18	0.03	0.28	0.08	0.09	0.24
0.28	0.00	-0.17	0.18	-0.17	0.18	0.14	0.00
0.35	0.44	0.23	0.39	0.48	0.36	0.11	0.46
0.39	-0.12	0.34	0.04	-0.04	0.14	-0.12	-0.27
0.36	0.58	-0.12	-0.03	-0.07	0.12	0.08	-0.03
0.28	0.58	-0.22	0.04	0.07	-0.04	0.08	0.07
0.39	0.51	0.31	0.08	0.22	0.37	0.09	0.10
0.11	0.48	0.38	0.02	0.07	0.26	0.00	0.00
0.15	0.18	0.36	0.04	0.36	0.20	0.08	-0.30
0.14	0.48	0.03	-0.03	0.03	0.25	0.11	-0.13
-0.12	0.43	0.32	0.38	-0.36	0.20	-0.17	-0.00
-0.01	0.19	-0.18	0.02	-0.04	0.04	0.08	-0.19
-0.04	0.48	-0.32	-0.08	0.08	-0.08	-0.12	0.01
0.38	matrix	0.12	0.08	0.06	0.17	0.14	0.27
0.38	0.48	0.19	0.02	0.02	0.04	0.20	0.17
0.34	0.81	-0.37	-0.06	0.26	0.03	-0.27	-0.17
0.38	0.38	0.18	0.11	0.48	0.00	matrix	-0.19
0.38	0.34	0.32	0.12	0.05	0.06	-0.20	0.30
-0.04	matrix	0.34	0.28	0.22	-0.01	-0.15	-0.07
-0.03	0.23	0.36	0.28	0.28	-0.07	-0.14	0.44
0.33	0.41	-0.28	-0.06	0.27	0.02	0.24	0.08
0.35	0.36	0.18	0.09	0.06	0.09	matrix	0.01
0.37	0.47	-0.18	0.30	0.18	0.08	0.14	matrix
-0.26	-0.28	0.34	0.11	0.27	0.00	-0.11	-0.25

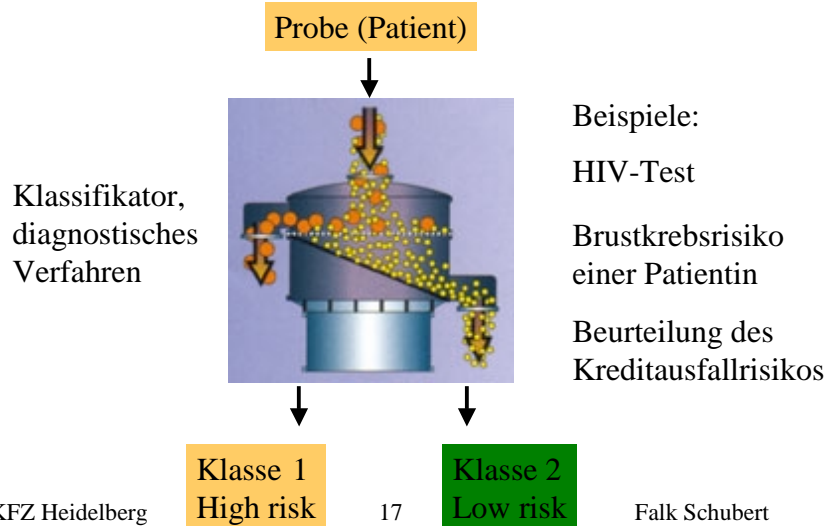
matrix-CGH

Aktuelle Forschungsziele

- Suche Kombinationen molekularer Marker zur Unterscheidung der Klassen hohes und niedriges Risiko
- Ranking der Marker für hohes Risiko
- Risikoabschätzung für individuelle Patienten



Zusammenfassung



DKFZ Heidelberg

Einordnung von Klassifikationsproblemen

Einordnung der Klassifikation

- Überwachtes Lernen
(supervised learning)
- Unüberwachtes Lernen
(unsupervised learning)

Einordnung der Klassifikation

- **Klassifikation:**
 - Vorhersage von Klassen (diskreten Größen)
- **Regression:**
 - Vorhersage kontinuierlicher Größen
 - Lineare Regression: $Y = \beta_0 + \beta_1 X$
 - Multiple Regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Klassifikation und Informatik

- Data Mining
 - Cluster-Analyse
 - Assoziationsregeln
 - Klassifikation
- Maschinelles Lernen (machine learning)
- Algorithmen basiert
 - schrittweises Verfahren zur Lösung /
Berechnung von Problemen

Klassifikation und Statistik

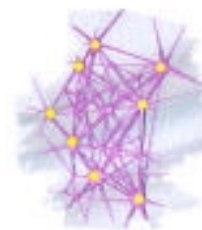
- Diskrimination
- Fehlerminimierung
- Mathematische Herleitung
- Verwendung von Regressionsproblemen

Klassifikation und Mustererkennung

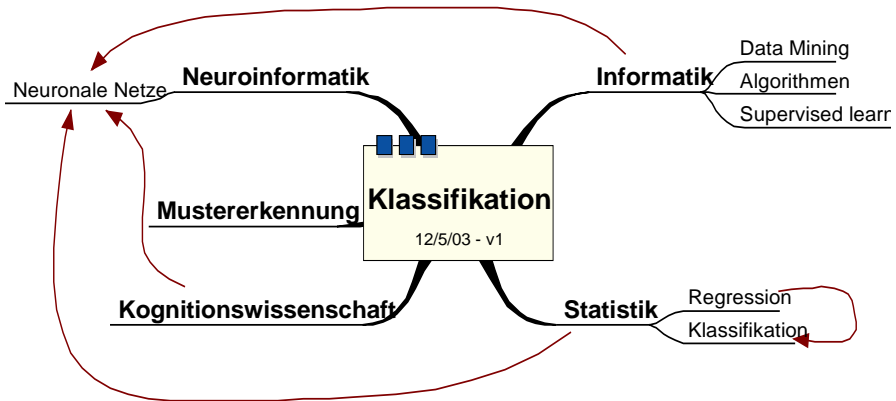
- **Muster**
 - Regularitäten
 - Modell
 - Konzept
 - bündelt Menge von Beobachtungen
- **Beispiele**
 - „Signatur“ von Tumoren
 - Apfelkonzept (rund, glatt, mit Stiel ...)

Klassifikation und Kognition

- **Prinzipien der biologischen Informationsverarbeitung**
- **Neuroinformatik**
- **Neuronale Netze**
- **Konnektionismus**



Zusammenfassung



DKFZ Heidelberg

25

Falk Schubert

DKFZ Heidelberg

Formale Definition eines Klassifikators

Merkmalsvektor, Zielvariable

Merkmalsvektor,
Attributvektor,
(feature vector)

Zielvariable
(class label)

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in X$$

$$y \in Y$$

Bei K Klassen meist

$$y \in [1, \dots, K]$$

Modellfunktion, Zielkriterium

Trainingsdaten

$$(x_1, y_1, \dots, x_N, y_N)$$

Zu lernendes Modell

$$\hat{f}(X)$$

Zielkriterium, Generalisierungsfehler

$$Err = E[L(Y, \hat{f}(X))]$$

Verlustfunktion

Zielkriterium, Generalisierungsfehler

$$Err = E[L(Y, \hat{f}(X))]$$

Verlustfunktion (loss function), Kostenfunktion (cost f.)

$$L(y_i, \hat{f}(x_i)) = \begin{cases} 0, & y_i = \hat{f}(x_i) \\ 1, & y_i \neq \hat{f}(x_i) \end{cases}$$

Beispiel: 0,1 Verlustfunktion

Hypothesenraum, Lernalgorithmus

Zu lernendes Modell

$$\hat{f}(X) \in \textit{Hypothesenraum}$$

Lernalgorithmus:

Algorithmus um den „optimalen“ Klassifikator aus dem Hypothesenraum für die Trainingsdaten zu suchen

Klassifikationsalgorithmus:

Lernalgorithmus + Hypothesenraum

Beispiel, Klassifikation von Äpfeln

Merkmalsvektor

Zielvariable
(class label)

$$x_i = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in X$$

$$y_i \in Y$$

$y_i = 1$, Objekt ist ein Apfel

$y_i = 2$, Objekt ist kein Apfel

x_1 Farbe, x_2 Größe, ...,

x_n Oberfläche

Beispiel, Klassifikation von Äpfeln

Trainingsdaten

$$(x_1, y_1, \dots, x_N, y_N)$$

Insgesamt N Beispiele für Objekte, die Äpfel sind bzw. keine Äpfel sind.

Zu jedem Objekt ist der Merkmalsvektor sowie das class label (Apfel, kein Apfel) angegeben.

Beispiel, Klassifikation von Äpfeln

Zu lernendes Modell

$$\hat{f}(X)$$

Funktion, die zu jedem Objekt aufgrund des Merkmalsvektors ermittelt, ob es ein Apfel ist oder nicht.

Die Funktion sollte so gewählt werden, dass möglichst alle Äpfel erkannt werden (Minimierung des Zielkriteriums).

Zusammenfassung

- Merkmalsvektor $x_i \in X$
- Zielvariable $y_i \in Y$
- Trainingsdaten $(x_1, y_1, \dots, x_N, y_N)$
- Zu lernendes Modell $\hat{f}(X)$
- Zielkriterium, Generalisierungsfehler
 $Err = E[L(Y, \hat{f}(X))]$
- Verlustfunktion
 $L(y_i, \hat{f}(x_i))$

Klassifikatoren

- Support vector machines (SVM)
- K-nächster-Nachbar (k-NN)
- Neural network (KNN, ANN)
- Entscheidungsbäume, Gruppen von Bäumen (forests)