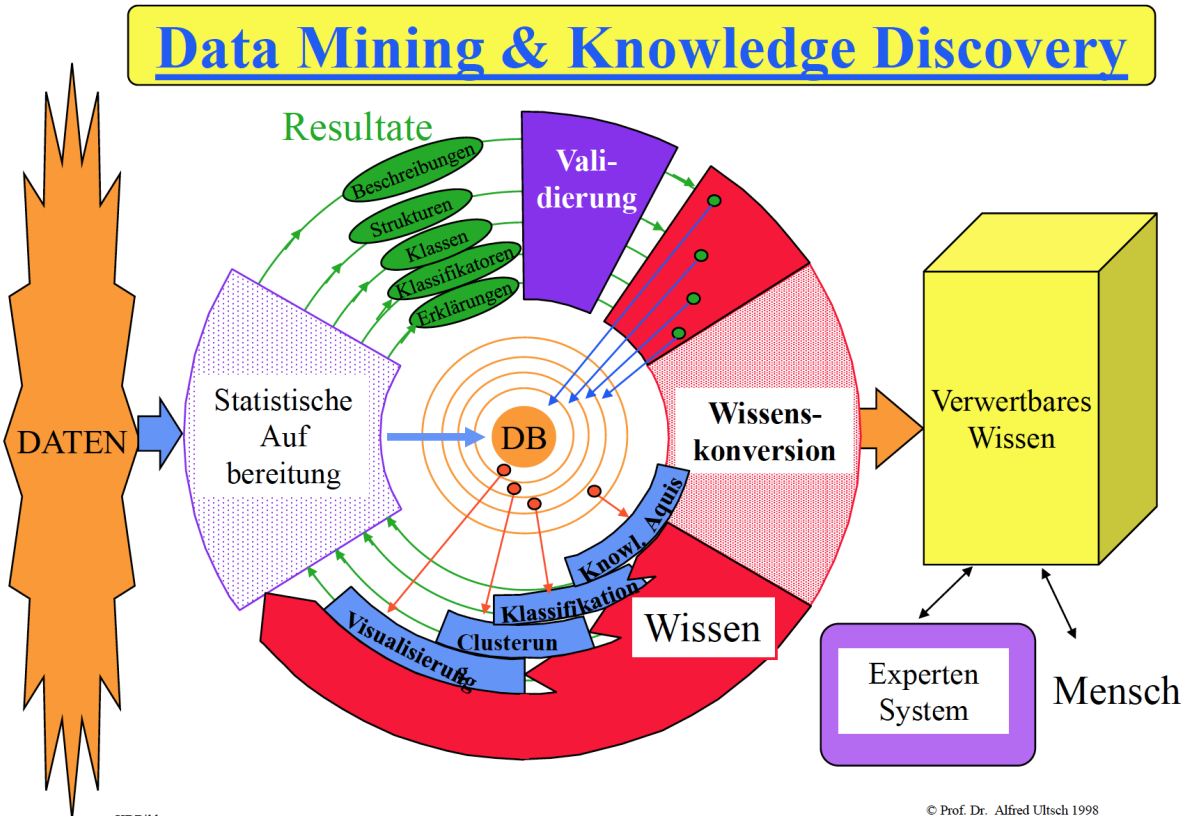
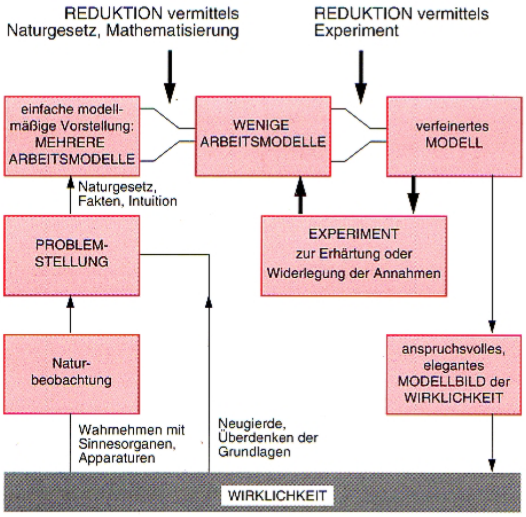
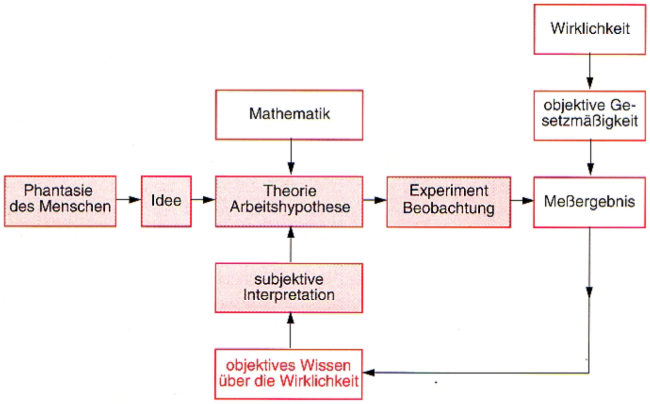
(a) Paradigma des Dataminings, in blau Methoden und in grün Gliederung



(b) Paradigma: Naturwissenschaften



(c) Paradigma der Physik



Statistik: gezielte Datensammlung bei Hypothese und Modell
Data-Mining: Erst aus Datensammlung folgt Wissen, kein Modell als Voraussetzung und keine Hypothese.
Physik: Beobachtung→Messung→mathematische Formulierung→Hypothese→Voraussagen →Experimentelle Bestätigung→Theorie→weitere Voraussage→keine Widersprüche zur Erfahrung→Naturgesetz. -
DB: Datenbank, Knowl.Aquis.: Knowledge Aquisiton - Wissenserzeugung und -darstellung. B/ Bäume, falls für Menschen verständlich, Expertensystem: Automatischer Beweiser, s. KI-Vl. (B/ Prolog)

Inhaltsverzeichnis

Kapitel 1

Beschreibungen

1.1. Statistik

Allgemein: Unter 10 Werte einer Variablen keine Statistik möglich, unter 100 Werte kein Data-Mining sinnvoll. Ein Datensatz entspricht der Zeile einer Tabelle, wenn eine Spalte einer Variablen zugeordnet ist. Grundlegende Begriffe:

- *Variable*: Datenspalte, auch Stichprobe.
- *Datensatz*: B/Merkmale einer Person, wie Geschlecht und Gewicht
- *Dichteverteilung*, auch pdf

Sinvolle Maße einer Verteilung hängen von Verteilungsform ab. Standardfall ist die Normalverteilung. Zentrale Momente der Statistik beschreiben Verteilungen folgendermaßen[1]:

1. Lagemaße: Mittelwert M, Median, Modus(Häufigkeitswert), Erwartungswert (QM)
2. Streumaße: Varianz, Standardabweichung SD, Spannweite
3. Schiefe v: (skewness) Maß der Abweichung von einer symmetrischen Verteilung,$v = 1$ bei Normalverteilung
4. Kurtosis: Exzess/Wölbung einer Verteilung Normalverteilung $w = 3$

Bemerkungen:

- Achtung: einige Moment sind nur sinnvoll bei einer Normalverteilung definiert,[2] z.B. SD oder M
- $v < 0$ linksschief, $v > 0$ rechtsschief, $v = 0$ symmetrisch
- robuste Kennwerte bei Ähnlichkeit zu Gauß: MAD, AMAD, stdrobust, etc.
- es existieren stat. Tests[3], die M oder SD oder w von Verteilungen vergleichen: Kruskall, T-Test, Wilcoxon, Anova etc.

1.2. Verteilung einer Variablen

Wichtigste Methoden zur Abschätzung der Verteilung einer Variablen:

- Histogramm: konstanter Kerndichteschätzer(Daten innerhalb von Hyperkugel in 1D), entspricht Bin, d.h. Teilintervall eines Wertebereiches, 2 Probleme: Binbreite und Startwert des Intervalls
- PDE-Graph: Dichtefunktion mit Pareto-Abschätzung eines variablen Kerndichteschätzer
- QQ-Plot: Plotte Werte gegenüber einer bekannten Verteilung in n Quantilen
- BoxPlot: Zeige Wertebereich, Ausreißer, 25, 50, 75 % Quartil

Achtung: Bei gebogenem QQplot ist der Boxplot vorsichtig zu interpretieren: Es werden unter Umstaenden zu viele Ausreißer angezeigt, s. **??**

[1] Formeln s. Wiki, Formelzeichen nach APA-Style
[2] bzw. müssen mit anderen Formeln berechnet werden
[3] APA-Notation t-Test: B/$t(54) = 5.43$ Testart(Freiheitsgrade 54)=Wert der t-Verteilung $p < .001$ Signifikanz

Vermutungen bei QQ-Plot:

- S-Kurve => Gleichverteilung
- 1 Richtung schief gebogen: quadratisch oder Wurzelfunktion
- beachte: über Transformationen zu verifizieren.

Unterscheide:

- ECDF - empirische cdf[4]): Daten werden durch Zufallsexperiment mit einer Wahrscheinlichkeit erzeugt, s. Gl. **??**
- Dichtefunktion - pdf[5], Likelihood: Ableitung der cdf (falls existiert), es gilt: Gl. **??**

$$ecdf(x) = \frac{Häufigkeit der Beobachtungen < x}{Häufigkeit aller Beobachtungen} \tag{1.1}$$

$$cdf(t) = \int_{-\infty}^{t} pdf(x)\ \mathrm{d}x \tag{1.2}$$

1.2.1. Wertebereiche und Ausreißer

Der BoxPlot zeigt zwar die Ausreißer explizit an, hat aber manchmal eine implizite Verteilungsannahme. Die Entscheidung, ob Ausreißer angezeigt wurden, ist schwierig zu treffen, weil nach Transformationen Ausreißer hinzukommen oder verschwinden können. Folglich muss erst eine richtige Verteilungshypothese gefunden werden.

- Wertebereich über BoxPlot, Ausreißer außerhalb der Whiskers, also der Ränder der Kastengrafik
- Suche nach Nans.
- Inputation, falls möglich, B/Über starke Korrelationen(Regression) oder Nullergänzung
- Statt relativer Änderung $R = \frac{B-A}{A}$, relative Differenz $D = \frac{B-A}{\neq frac12(B+A)}$ verwenden.

Abschätzung der Genauigkeit:

- kleinste messbare Differenz ist Null => doppelte Werte
- kleinste messbare Differenz ist nicht Null => Genauigkeit

Achtung: SD ist schwerer zu schätzen als M. Dies ist die kritische Größe bei der Definition eines gültigen Distanz (**??**), insbesondere sehr Anfällig für Ausreißer.

1.2.2. Multimodale Verteilungen

Mehrere Moden sind ein Hinweis auf eine mögliche Gruppenbildung der Daten. Sollten Moden in Daten vorher erkennbar sein oder nach einer Transformation erkennbar werden, ist es möglich Gruppen zu definieren. In einer Variablen, welche nicht normalverteilt ist, ist dies mit leichtverständlichen Ansätzen nur heuristisch möglich.

Bei normal verteilten Variablen wird das Gaußmixturen Model (*GMM*) verwendet. Für die Abschätzungen der Mittelwerte und Standardabweichungen eignet sich der Erwartungs-Maximierung-Algorithmus (EM). Der EM-Algorithmus sucht ein lokales Maximum dreier Parameter jeder Mode. Er benötigt eine vorgegebene Anzahl an Moden und die jeweiligen Mittelwerte, Standardabweichungen und Gewichtungen. Mithilfe des Bayes-Theorems konvergiert der Algorithmus iterativ. Dabei Minimiert er lokal den Fehler zwischen GMM Annahme und Dichtefunktion. Über Bayes können dann empirisch Grenzen zwischen den Moden berechnet und somit den Daten Klassen zugeordnet werden.

[4] cumulative distribution function
[5] probability density function

Begriffe:

- *A-priori*: Anfangswahrscheinlichkeit der Verteilungsfunktionen P(X=x), P(G=i)
- klassenbedingte Wahrscheinlichkeit: $P(X = x|G = g)$ Unter der Voraussetzung der Existenz der Klasse g, das Vorkommen der Daten x, auch Likelihood
- *Posteriori*: $P(G = g[X = x) = p(g|x))$, ist die Wahrscheinlichkeit der Zugehörigkeit eines Datensatzes zu einer Klasse.

Satz von Bayes[6]

$$P(g|x) = \frac{P(g) * p(x|g)}{N} = \frac{priori * bedingte\ Wahrscheinlichkeit}{Normalisierung} \tag{1.3}$$

1.3. mehrere Variablen

Betrachtungsarten:

- synoptische Gesamtschau:B/ Heatmap, zirkulare Pixel
- Scatterplots: 2Dimensionale oder 3 Dimensionale Streudiagramme
- Vorgriff: Für Scatterplots werde durch Transformationen beeinflusst
- Leiterdiagramme: Nachteil ist Reihenfolge der Variablen
- Andreus Kurven

Tipp: Figuren/Strukturen in Scatterplots: Daten besitzen einen Zusammenhang, aber beachte, dass sie nicht unbedingt korreliert[7] sein müssen.

1.3.1. Zusammenhänge und Korrelationen

Ziel: Entkorrelation der Variablen durch Auschluss bei hoher Korrelation oder Entkorrelierung. [8]. **Begründung:** Bei Korrelierten Variablen wird die Distanz dieser Variablen im Bezug zu anderen Variablen verhältnismäßig überproportional gewichtet. Zwei Grundtypen

1. Pearson: Flächenvergleich, parametrisch → nur bei Normalverteilten Variablen erlaubt und **linear**. Braucht als Voraussetzung die Normalverteilung - **Korrelationskoeefizient**
2. Spearman: Sortiere Daten nach Rank, d.h. Rankvergleich der Fläche→nicht parametrisch (keine Verteilungsannahme), aber immer noch **linear - Korrelationskoeefizient**
3. Kendalls Tau: Suche nach Ordnungen, Differenz der Wahrscheinlichkeit → **nicht linear - Statistischer Zusammenhang**
4. mehrere Variablen: *Kovarianzmatrix*

Bemerkung zu Spearman: „Bei Vorliegen von metrischen Daten sollte im Allgemeinen nicht auf einen Rangkorrelationskoeffizienten zurückgegriffen werden, da mit der Rangvergabe auch ein beachtlicher Informations¬verlust einhergeht.“

- Keine Korrelation bedeutet nicht, das Variablen keinen Zusammenhang miteinander haben. B/ Zwei sich kreuzende Linien im Scatterplot
- Aber Korrelationen müssen nicht unbedingt auf kausale Zusammenhänge hindeuten. B/ Zwei voneinander getrennte Punktwolken
- Von den drei Korrelationen liegen zwei Maßen (Pearson und Spearman) ein Modell zu Grunde

[6]Oft ist $N = P(X)$ und dies unbekannt
[7]immer linearer Zusammenhang in Scatter, d.h. Regressionsgerade einfach einzeichbar
[8]z.B. über PCA oder Mahalanobis

Nach Berechnung der Kennzahl berechne über Permutationsexperiment (n-fach) die Verteilung des Korrelationskoeffizienten und vergleiche die Fläche rechts vom echten Wert der Korrelation im ecdf-Plot. Signifikant[9] falls $p(Fläche) < 0.05\%$
Achtung: In richtigen Datensaetzen kommen Fehlstellen (NaN) oft vor! Diese muessen vor einer Berechnung des Korrelationskoeffizienten ersetzt werden.

- Ersetzung durch Null, o.ä., führt zu Fehlern
- Spearman/Pearson: Einzelbeiträge werden am Mittelwert gemessen => Zuweisung der fehlenden Daten zu den jeweiligen Mittelwerten ergibt keinen Einfluss der NaNs
- Kendall: schwierig, plausible vorgeschaltete Substitution nötig (z.B. KNN-Classifier (Vorgriff auf später))

1.4. Vorgehensweise bei einer ersten Inspektion

1. synoptische Gesamtschau mit prozentuale Skalierung[10].
   - Gegenüberdarstellung aller Variablen mit BoxPlots
   - Bestimmung der Anzahl an NaNs: Inputation der Nans, falls möglich, bei keinen NaNs Vorsicht!
   - Abschätzung der Genauigkeit
2. Verteilungsbetrachtungen einzelner Variablen: Ziel ist Abschätzung der **Varianz**, der Quantisierung, vielleicht sogar Verteilungsart
3. Zusammenhänge:
   - Korrelationen über *Kovarianzmatrix*: Ziel ist Entkorrelierung/Auschluss von Variablen
   - Scatterplots: Strukturbetrachtung
   - Gegenbenenfalls: Hat Vorklassifikation[11] sichtbare Auswirkungen?
     B/ männlich $<->$weiblich

1.5. Transformation

Es soll eine Transformation gefunden werden, durch welche die Daten standardisiert werden können. Diese ändert immer das Verhältnis der Daten untereinander. Falls eine einmodige Normalverteilung vorliegt, ist dies die *Z-Transformation*. Die Normalverteilung ist am Besten im QQ-Plot als Gerade sichtbar. Begriffe:

- *Z-Transformation*: $z = \frac{x-M(x)}{SD(X)}$, Kritisch: Definition des Streumaßes, da Standardabweichung oft nicht plausible Verteilungsbeschreibung. Achtung: Verlust der Skalierungsverhältnisse
- *Mode*: Anzahl nur heuristisch bestimmbar, Argumentation über Ockhams Rasiermesser
- Über cdftrans auch die Transformation auf eine Gleichverteilung möglich

Ziel: Transformiere Daten plausibel zu einer Gaußverteilung, oder s.d. ein einigermaßen vernünftiger Wertebereich entsteht[12]. **Dadurch werden die Lage und Streumaße vergleichbar.**
Für die Suche nach einer geeigneten nicht linearen Transformation ist der BoxCox-Algorithmus hilfreich. Mithilfe der Maximum-Likelihood-Schätzung[13] wird ein Wert λ für eine Transformation nach

[9] Notation nach APA: $r(80) = .98$ corr(Freiheitsgrade-2)=Stärke, $p < .05$
[10] u.U. vorher nach NaNs oder Rang einer Variablen ordnen
[11] Datensätze einfärben/getrennt zeichnen
[12] große Streuung im Scatterplot, Dichteverteilungen für alle Variablen gleich
[13] Algorithmus auch gut bei, falls a priori für *GMM*-Trennung nicht bekannt

Gl. **??** berechnet. Dieser Wert liefert eine Schätzung der notwendigen Verbiegungskraft, um nach Transformation eine Gaußverteilung zu erreichen.

$$x- > \frac{x^{\lambda}-1}{\lambda} \tag{1.4}$$

Dann kann über die *Potenzleiter* eine plausible Transformation[14] gewählt werden.

1. Box Cox Algorithmus ergibt eine Vermutung über Verteilung
2. *Potenzleiter* zur Zuordnung des λ benutzen, und mit Wissen Vermutung bestätigen.
3. Bei logarithmieren versuchen log10 statt ln zu nehmen da verständlicher, aber der Unterschied nur in einer multiplikativen Konstate entspricht
4. Beachte bei logarithmieren: Wertebereich kleiner 8 ergibt einen großen Fehler bei einer slog=log(x+1) Transformation => Wertebereich zuerst verschieben, prozentuieren oder mit einer Konstante multiplizieren

Tipp: Nach geeignet Transformation sollten Daten im Streudiagramm schön weit verteilt sein.

1.6. Definition von Ähnlichkeiten und Distanzen

Die Vorverarbeitung der Daten erlaubt eine Bestimmung einer gültigen Distanzmaßes oder die Definition eines gültigen Ähnlichkeitsmaßes[15]. Nur Distanzmaße können in Projektionsverfahren verwendet werden. Da Dimension eines Datensatzes über die Anzahl der Variablen m definiert ist, wird die Distanz jeweils paarweise zweier Zeilen über alle Variablen berechnet. Sei die Zeilenanzahl d so bildet sich daraus die Distanzmatrix (d^2, d^2) D.

Die Umrechnung zwischen Distanzmaß D und Ähnlichkeitsmaß A erfolgt im symmetrischen Fall mit

$$D(i,j) == max(A) - A(i,j) \tag{1.5}$$

Und die Rücktrafo:

$$A(i,j) = \frac{1 - D(i,j)}{max(D)} \tag{1.6}$$

Im Unsymmetrischen Fall ergibt sich je paarweise Distanz $d(i,j)$ und Ähnlichkeit $s(i,j)$

$$d(i,j) = \sqrt{s(i,i) + s(j,j) - 2s(i,j)} \tag{1.7}$$

Die Definition einer Distanz ist erst dann erlaubt, wenn alle Variablen ähnliche (innere[16]) Varianzen haben!. Deswegen sind die oberen Abschnitte von entscheidender Bedeutung. Ein Trick für die Definition der euklidischen Distanz ist der folgende: Transformiere, s.d. innere Varianzen der Variablen <1 und Varianzen zwischen den Variablen >1. Damit nutzt an die Skaleninvarianz der Euklid-Metrik aus.

Ein gute Visualisierung ist die *U-Matrix*. Für die Praxis bedeutsame Eingenschaften einer Distanz sind:

1. Skaleninvarianz
2. Rotationsinvarianz
3. Translationsinvarianz
4. segment additiv

Bei vielen Variablen schaue durch Ähnlichkeitsmaße ihre Verteilung im Hochdimensionalen an und verwende dabei Vorteile, wie Rotationsinvarianz und Entkorrelierung. Arten von Distanzen:

[14] $B/\lambda \to 0 => log$
[15] In der Praxis kostet dies 90 % der zur Analyse benötigten Zeit (eines Menschen)
[16] Die Varianz innerhalb einer Variablen ist vergleichbar mit inneren Varianz einer anderen Variablen

- Metriken
- Normen[17]
- *Ultrametrik*
- *geodätische Distanz*

Typen von Metriken:

- Minkowski: kritischer Punkt: d=1
  - Nicht skaleninvariant, aber Euklid (p=2) ist rotationsinvarant
  - Pathologie: d<1 gestaucht, und d>1 gestreckt
- Mahalanobis Distanz: rechnet Korrelationen weg bzw. unkorrelierte Variablen werden korreliert; Berechnet sich aus inverser Kovarianzmatrix, **Vorteil: Entkorrelierung**
- Hemming-Distanz bei Ordinal-Skala

Bem.: hochkorrelierte Daten bedeuten eine Gewichtung der Daten in den Metriken **Hauptziel**:

1. Definierung korrekter Distanz (**kritisch!**), über geeignete Transformation der Form, dass
   - d<1 für nicht interessante Distanzen
   - d>1 für interessante Distanzen
2. Validierung über Experten.

1.6.1. Datenreduktion

Aus rechnerischen Gründen ist die Berechnung einer Distanzmatrix größer als 3000x3000 langwierig. Deswegen nimmt man ein geschickt gewähltes Sample der Daten.

- Keine Informationen über den Datensatz: Zufalls-Sample
- Vorklassifikation gegeben: Sample Größe bezüglich jeder Klasse getrennt -> Klassenweise sampeln
- Oft kann angenommen, daß der 80% bzw. Großteil der Daten gewöhnlich sind, also keine neuen Informationen ergeben.

Vorgriff auf Clusterung: Nach dem Schritt der Clusterung des Samples ist folgendes möglich. Anschließend wird auf den ganzen Datensatz extrapoliert: z.B. wird ein KNN Klassifizierer mit der oben vordefinierten Distanz gewählt mit k=3 oder k=5 Nachbahrn, damit für jeden Datensatz eine Mehrheitsentscheidung erzwungen wird: Für jeden Datensatz werden als Nachbahrn schon vorklassifizierte Datensätze gesucht bei unterschiedlicher Klassenzuordnung der Nachbahrn erfolgt ein Mehrheitsentscheidung.

1.7. hochdimensionale Daten

Grundsätzlich sind dichtebasierende und abstandsbasierende Datensätze zu unterscheiden. Sei $n \to \infty$, =>

1. Volumen der Hyperkugel direkt auf Schale unter Oberfläche verteilt => Daten liegen auf der Schale unter der Oberfläche der Hyperkugel
2. $max(d(x,y)) - min(d(x,y)) - > 0$, d.h. Varianz wird kleiner
3. Zur Analyse idealerweise 100^n Daten nötig

Folgerung: R^n ist in der Regel leer=> (**Hoffnung!**) Daten liegen auf Untermannigfaltigkeit U
Falls Nachbarschaftstreue des Projektionsraumes vorliegt **Ziel**: Abschätzung von $dim(U)$

[17]Metrik mit definierten Nullpunkt und linearer Struktur des Vektorraums, Kugeln sind konvex

1.7.1. intrinsische Dimension

auch Hausdorff Dimension ist die minimale Dimension des Unterraumes (Mannigfaltigkeit) von deim Eingaberaum, in dem die Daten liegen. Berechnung der intrinsischen Dimension über:

- VC-Dimensionen[18]: f(VC) ist definiert als $M = x \in X$, die diese Funktion in allgemeiner Lage auf dem Eingangsraum separieren kann.
- Fraktale Dimensionen: Einteilung des Eingaberaumes in gleich große Hyperwürfel
- Grassberger-Procaccia

Abschätzung der Untermannigfaltigkeit als B/ Grassberger-Procaccia

1. Korrelationsdistanzen C(r): Distanzen kleiner als Radius der Hyperkugel r
2. log(C(r)) gegen log(r) auftragen
3. ergibt Gerade => Steigung ist intrinsische Dimension

Achtung: die Methode versagt bei gruppierten Daten.

1.7.2. Topologieerhaltung

Unter der Annahme das Topologieerhaltung Nachbarschaftserhaltung bedeutet, besteht die Hoffnung eine sinnvole Untermannigfaltigkeit des hochdimensionalen Raumes zu finden, dabei entstehen folgende drei Probleme:

1. zu wenige Daten
2. Projektionen auf 3D schwierig wegen Informationsverlust
3. kleinste und größte Distanz werden immer ähnlicher, genauer:
   - $euklid = \sqrt{\sum_i \Delta_i^2} \underrightarrow{n\ gross}$ Gaußförmig $\underrightarrow{n \to \infty}$ Dirac Stoß („Peak“)

[18] Vapnik-Chervonenkis

Kapitel 2

Strukturen

Sei eine Projektion[1] als Matrix aus Datensätzen(Zeilen) und Variablen(Spalten) definiert. So wird allgemein zwischen linearen und nicht lineare Projektionen unterschieden. Dabei sind lineare Projektionen nur Rotationen. Im folgenden werden PCA, ICA und Projection Pursuit vorgestellt.

Bei nicht linearen Projektionen wird die Matrize verändert. Diese Projektionen sind dadurch generell fehlerbehaftet, da eine Stauchung/Streckung des Raumes auf eine Untermannigfaltigkeit stattfindet, welche nur gewissen Merkmale der Wirklichkeit entspricht[2]. Alle anderen noch folgenden Verfahren sind, wenn nicht anders erwähnt, nicht linear.
Achtung: Finden Verfahren Strukturen, so sind diese immer schwer zu verifizieren!
Ein Projektionsverfahren findet Strukturen => noch keine Aussagekraft => Zur Überprüfung immer mehrere Projektionsverfahren verwenden!

2.0.3. PCA

PCA makes the assumption, that the directions of input space with high variance have the most information about the dataset [Hotelling, 1933]. The coordinate system of the input space is replaced by a (principal) coordinate system, where the variance of data is maximized. It is done by finding a set of weighted linear composites of original variables, where the weights are found by eigen-decomposition. An equivalent definition through an objective function was defined by [Pearson 1901], where the average projection cost was minimized. The projection cost is defined by the mean squared distances between points ... The minimization of J is solved by choosing the basis vectors as eigenvectors of the covariance matrix with the constraints of orthonormality conditions [Duda et al 2001, p.115ff]. ... Now, the objective function E can be redefined through the eigenvalues as .. where n is the dimension of the input space I, m ist the dimension of the output space O. The largest eigenvalues correspond to the 1..m dimensions with the largest variance. Input space dimension with a small variance are discarded. Thus, PCA is an orthogonal projection of the data into a space of lower dimension.

Principal Component Analysis - ein linearer Projektionsverfahren

- Annahme: Erklärungswert ist die Varianzgröße
- Ziel: Korrelationen entfernen
- Hauptkomponentenanalyse: $\vec{y} = R\vec{x}$ mit Rotationsmatrix R, rotiere so, dass 1.Achse auf größter Varianz liegt und der Ursprung liegt auf dem physikalischen Schwerpunkt. Suche also orthogonale Achsen mit größter Varianz absteigender Reihenfolge
- Der PCA-Fehler wird minimal <=> Eigenvektoren der Kovarianzmatrix werden verwendet.
  - die Eigenwerte entsprechen den Varianzen
  - Diese Faktoren[3] strecken oder stauchen die Datenpunkte
  - Ein weiterer Output sind Faktorwerte d.h. die Matrix aus rotierten Datensätzen und Variablen
- Gruppierungen und PCA schließen einander aus
- Scree-Plot: Faktor-Anzahl finden, der die meiste Varianz beinhaltet

[1] Häufigkeit der Verwendung von Projektionsverfahren: MDS und PCA 90%
[2] B/ Zerdrücken einer Dose, 3D→2D
[3] Basisvektoren im Sinne von Variablen, coefficients/loadings,

Bessere PCA ist die Karhunen-Loeve Transformation[4] spezielle Auswahl von Eigenwerten $\lambda > 1$.
Fazit: Runterskalierung auf m=n-l Dimensionen, in denen die meiste Varianz vorhanden ist

2.0.4. ICA

"Independent component analysis (ICA) is a method for finding underlying factors or components from multivariate (multi-dimensional) statistical data. What distinguishes ICA from other methods is that it looks for components that are both statistically independent, and nonGaussian" [Hyvärinen et all 2001].

Independent Component Analysis - ein linearer Projektionsverfahren

- Annahme: Gaußverteilung bedeutet Zufälligkeit, also weißes Rauschen
- Suche nach Dimensionen, die möglichst nicht normalverteilt sind, B/ über Kurtosis (s. **??**)
- Annahme: Unabhängigkeit der Variablen, Voraussetzung ist Unkorreliertheit
- Tipp: Wende PCA vor ICA an, dann Variablen zumindestens unkorreliert - „Whitening"
- Für alle Variablen gilt: rotiere Variablen, s.d. sinnvollste gefunden werden ->Gruppensuche
- Algorithmus: Maximierung der Negentropie[5], relevant hier: geeignete Kontrastfunktion) kann gewählt werden

2.0.5. Projection Pursuit

ein linearer Projektionsverfahren

- Verlangt von Nutzer Kriterium für Minimierung, genannt Indexfunktion
- Oft mit Kernelfunktion verwendet $K(r) = \frac{1}{2*\pi} e^{-\frac{r}{2}}$
- Kernelfunktion: mittlere Distanz zum nächsten Nachbarn

2.1. Lyapunov

Basiert auf dem Fehler des Abstandes, genauer der Minimierung einer Fehlerfunktion. In Gl. **??** ist die allgemeine Energie- bzw. Fehlerfunktion gezeigt. Diese punkt-paarweise Distanzdifferenz zwischen Eingabe- und Ausgaberaum wird über die Minimierung eines lokalen Gradienten verglichen (*Gradient-Abstiegsverfahren*).

$$E(E, A) = \sum_{i,j} (E(i,j) - A(i,j))^2 \tag{2.1}$$

2.1.1. MDS

Multidimensional Scaling (MDS, [Kruskal 1969] tries to preserve pairwise distances D(l,j) as well as possible. Originally it was published by [Torgeson 1952] Therfore, MDS minimizes the objective (error) function E defined with ... where f(D(l,j)) is a nonmetric, monotone transformation of the distances in the Input space [Kruskal 1969, p.7]. E is often called stress and the minimization of E attempts to reproduce the general rank-ordering of distances. However, the error function E depends on the scale in which the distances are measured. It is preferable to normalize the error function E in order to reduce it to the same units as the distances. Sammons Mapping [Sammons 1969] uses therefore the error function ... The minimization is usually performed by gradient descent.

[4] weiterer Output: Faktorladung entspricht der Korrelation zwischen einem Merkmal und einem Faktor

[5] Entropiedifferenz zu einer entsprechenden normalverteilten Zufallsvariable

Multidimensionale Skalierung[6]

- es gehen nur Distanzen zwischen Eingaberaum E und Ausgaberaum A ein
- Tipp: überprüfe Güte mit *Shepard Diagramm*:
  - Gut, wenn ca. nur Diagonale
  - mäßig => starke Korrelation
- Qualität der Abbildung auch über Shephard-Kruskal Stressmaß
- Nachteil: Für Ausreißer wird versucht gute Darstellung zu finden, d.h. große Distanzen werden bevorzugt
- Nähe von Datenpunkten nicht klar definiert, aber Klassen über große Distanzen definierbar => Verdichtungen bleiben unklar

Trick: **MDS**: Transformiert Distanzen zu Punkte zurück.

2.1.2. Sammons Mapping

Bem.: **MDS** gut für große Distanzen, aber dort auch große Fehler durch Ausreißer möglich => große Fluktuationen im Shepard-Diagramm
Trick: Deswegen Normiere durch die Dimension des Eingaberaums

2.2. Graphen basierende Verfahren

Es gibt prinzipiel drei verschiedene Arten von Graphen

- *KNN*-Graph: für jeden Punkt k-nächste Nachbarn
- R-Kugel Graph: Für jeden Punkt, Verbindungen zu allen Punkten innerhalb von R-Kugel
- vollständiger Euklid-Graph

Ausgangspunkt der folgenden Graphen ist der Euklid-Graph, der alle Punkte als Knoten und alle Verbindungen zwischen allen Punkten gewichtet mit den Distanzen enthält. Er wird als gewichtete D-*Adjazenzmatrix* dargestellt. In der Regel verwendet man Teilgraphen, sogenannte Nachbarschaftsgraphen:

2.2.1. Teilmengen des Euklid-Graphen

Folgende Graphen sind echte Teilmengen voneinander:
vollständiger Euklid-Graph→Delaunay Graph→Gabriel Graph→RNG→MST
Definitionen:

1. Delaunay Graph: Ausgangspunkt ist *Voronoi-Zelle*→Darstellung benachbarter Zellen
2. Gabriel Graph[7] Definiert über Kreis mit Radius um Datenpunkt, wobei beide Punkte am Rand liegen → in diesem Kreis darf kein anderer Punkt liegen => Verallgemeinerung bei dim(n) sind Hyperkugeln
3. RNG: relative neighbour Graph: Über zwei Kreise um jeweils 2 Punkte ist die Schnittmenge definiert, in welcher kein weiterer Punkt liegen darf
4. MST: minimum spanning tree („Spannbaum“)
   - Baum bedeutet: Alle Punkte sind genau einmal miteinander verbunden, wobei der Abstand minimiert wird
   - k-1 längste Kanten ausschneiden → Cluster

[6] bekanntestes Beispiel für Abstandsfehler Verfahren
[7] „geradlinig benachbarte Voronoi-Zellen“

2.2.2. ISOMAP

[8]

- Approximierung einer *geodätischen Distanz*.
- Ausführung von MDS mit dieser Distanzdefinition
- **Knackpunkt**: wähle geeignetes k
- Im Idealfall Untermannigfaltigkeit von geodätischen Distanzen

2.2.3. LLE

local linear embedding: lokale PCA wird linearisiert und über *KNN* gewichtet. Dabei wird ein Datenpunkt aus der Linearkombination von Datenpunkten rekonstruiert: Distanzen der Approximation. Auch lokale lineare Interpolation
Knackpunkt: k ist hier beliebig, aber k>2 und k<n, n=?

[8] Datenbionik AG

2.3. Datenbionische Projektionen

Grundidee ist immer aus der Natur übernommen und wird dann als Algorithmus verfasst. Ein Prinzip sind **unüberwachte Lernverfahren** (s. **??**). Ihr Funktionsprinzip beruht auf der biologischen Erkenntnis, dass viele Strukturen im Gehirn eine lineare oder planare Topologie aufweisen. Die Signale des Eingangsraums, z. B. visuelle Reize, sind jedoch multidimensional. Es stellt sich also die Frage, wie diese multidimensionalen Eindrücke durch planare Strukturen verarbeitet werden. Biologische Untersuchungen zeigen, dass die Eingangssignale so abgebildet werden, dass ähnliche Reize nahe beieinander liegen. **s. Neurophysik:** Somatotopic Karte: Homunculus.
In VL nicht behandelt wurde MLP-BP, da „schlecht“

2.3.1. ESOM

Eine **self-organizing map** wird oft auch Kohonenkarte genannt. Mit dieser selbstorganisierenden Karte bezeichnet man eine Art von künstlichen neuronalen Netzen, wobei multidimensionale Daten auf den zweidimensionalen Raum analog zur somatotopischen Karte runterprojeziert werden[9].
***Emergent*e SOM**: Projektion auf Gitter aus Neuronen/*Units*. Zuerst wird eine Nachbarschaftsdistanz definiert. Aus den Eingabedaten wird die *Best Matching Unit*(BMU) ausgewählt. Durch „Lernen“ mit einer vordefinierten *Lernrate* wird nun durch die Nachbarschaftsfunktion[10] die Gewichtung der Vektoren angepasst. Danach wird die Nachbarschaft durch die Lernrate verkleinert[11].

- Durch Lernrate konvergiert Algorithmus
- *Units* in Gitterform geordnet, beachte:
  - rechteckig wie hexagonal gleiche Leistung, aber quadratisch schlechter
  - Randeffekte müssen berücksichtigt werden => besser: Torus, also über Gitterverbiegung kein Rand (s. QM2, Festkörper)
- ESOM ist problematisch, falls Daten dichtebasierend sind
- Für ESOM relevante Parameter:
  - Anzahl an Units
  - Anzahl an *Epoche*n, d.h. wie lange soll gelernt werden?

Die Visualisierung der ESOM erfolgt über die U-Matrix.
Praxis:

- BMUS sehr nahe beieinander und auf Hügeln => ESOM untertrainiert/zu wenige Epochen
- BMUS sehr extre gleichmäßig verteilt, Eierhalenform, d.h. jeder Punkt in einem See umrandet von Hügeln => ESOM übertrainiert
- Punkte gleichmäßig auf Rändern verteilt =>Randeffekte => toroides Gitter wählen

[9] SOM: Beweis das Minimierung existiert
[10] Gauß, Kegel, etc
[11] out(linear), exponentiell, lead in (Tiefpass)

2.3.2. U-Matrix: Visualisierung der ESOM

Umgebungsmatrix: Höhendarstellung lokal integrierter Distanzen, also der Nachbarschaft. Aus großer Höhe folgt viel Platz im Eingaberaum und dadurch große Abstände. Es sind also gut dimensionale Brüche zu sehen. In Ebenen sind ähnliche Datensätze gesammelt.

Dabei sind Faltungen Strukturen, die durch U-Matrix nicht gut gezeigt werden können.
(Bei Kombination mit SOP: Die Höhe der U-Matrix ist unabhängig ob Mittelwert gezogen wird oder nur summiert wird, aber die Gewichtung des Punktes wird mit den Punkten um die Daten herum verglichen.)
Tipp: „Vulkan“ mit einzelnen Punkt ist wahrscheinlich ein Ausreißer.
Self organizing maps (SOM) where invented by [Kohonen 1984]. To exploit emergent phenomena in SOMs [Ultsch 1999] argued to use a large number of neurons (>4000). By gaining the property of emergence through self-organization, this enhancement of SOM is called emergent SOM (ESOM): Let M be a set of neurons (map positions) with the corresponding set W weights, then the SOM training algorithm constructs a nonlinear and topology preserving mapping of the input space by finding for each l∈I the bestmatching unit ... In each step the SOM learning is achieved by modifying the weights in an neighborhood with ...

The cooling scheme is defined by the neighborhood function ... and the learning rate ..., where the Radius R declines until R=1 through the definition of the maximum number of epochs. Note, that there exists no objective function for ESOM. The topology of the feature map is toroid, i.e. the borders of the map are cyclically connected [Ultsch 1999]. The positions ... of the bestmatching units exhibit no structure in the input space [Ultsch 1999]. Only with the exploitation of a visualization technique for SOMs called Umatrix [Ultsch, Siemon 1990], the structure of the input-data emerges. The accerage of all data distances of a neurons weight vector w is called U height:

where N(i) are the immediate neighbors of .... A display of all U-heights is called U-Matrix[Ultsch, Siemon 1990]. The U-Matrix is an approximation of the abstract U-matrix, which formalizes the structures of a U-matrix such that the height structures of a U-matrix are defined by Voronoi borders of the points in the output space[Lötsch, Ultsch 2014].

abstrakte U-Matrix

Verallgemeinerung einer Visualisierung von beliebigen Projektionen über Voronoi-Zellen. Beispiel: Zwei Radien um zwei Punkte, Linie durch Schnittpunkte, Abstand der Punkte zur Linie ergibt die Höhe der Distanzen in R^n
Praxis: SOM & *Best Matching Unit* nicht hüpfen lassen => abstrakte U-Matrix

2.3.3. ABC und Schelling

Das **Schelling Modell**, ein Segregationsmodell: Bei noch so geringerer statistischer Präferenz neigen benachbarte Punkte sich nach Ähnlichkeiten zu ordnen, falls sie die Möglichkeit zum Hüpfen bekommen. Dadurch entsteht eine Clusterung.
Ant based Clustering: Produziert viele kleine Cluster, da Maximierung von Topographie*Ausgabedichte

2.3.4. SOP

swarm organized projections[12]:
SOP= Schelling-Modell+*KNN*+1Ant(*Databot*) & Fokusierend

- Besitzt keine Funktion der Min/Maximierung
- nur gut in Kombination mit U-Matrix
- **Vorteil**: gut für massiv nicht lineare Daten, da kein Modell zugrunde liegt[13]

[12]Verfahren der AG Datenbionik
[13]z.B. Lyapunov Funktion, Formel, Geometrie. etc.

2.4. Fokussierend Abbildungen

Verfahren, welche ähnlich wie SOM funktionieren. Es werden also erst globale Strukturen betrachtet und dann wird auf lokale Strukturen eingegangen, indem der Nachbarschaftsradius von groß zu klein skaliert wird,

2.4.1. CCA

When unfolding nonlinear structure, MDS cannot reproduce all distances. Therefore, [Demartines/Herault 1995] proposed a projection method favoring local neighborhoods. CCA tries to reproduce short distances before reproducing long distances later. The error objective is defined with ...

Curvilinear Component Analyses

- **CCA=SOM+MDS**
- Projektionspunkte werden zufällig bestimmt
- Minimiert topographischen Fehler

2.4.2. t-SNE

The t-distributed Stochastic Neighbor Embedding (t-SNE) is an enhancement of SNE [Hinton and Roweis, 2002], where the Kullback-Leibler divergence is symmetrized amd the crowding problem solved. The latter is done by redefining the conditional probabilities of the output space O by an application of a Student t-distribution with ... In [Van der Maaten and Hinton 2008] the distance between datapoints is redefined as the conditional probability that j would pick l, l, ... is the variance of the Gaussian that is centerend on the data point j. If the projection is correct, the conditional probabilities will be equal [Van der Maaten and Hinton 2008]. Therefore, the objective function is defined by the symmetric Kullback-Leibler divergence with ...

stochastic neighbour embedding. Mit $S_i = S_i(k)$ als Nachbarschaftsbreite, $p = p(x_i|x_j)$ und q sind Dichtefunktionen
SNE:

- Fehlerfunktion ist *Kullback Leibler Divergenz* mit $KLD = \int plog(\frac{p}{q})$
- „crowding problem“[14], KLD ist unsymmetrisch
- Gaussverteilung

t- SNE:

- $SKLD = \sum plog(\frac{p}{q})$
- t-Verteilung

Problem: Wahl der Nachbarschaftsbreite S

2.4.3. GTM

GTM: Generative Topographic Mapping: Lyapunov+GMM mit EM+Kette (Nachbarschaft), ist dichtebasierend im Eingaberaum

[14] s. Wiki/Google

Kapitel 3

Clusterung

Clustern bedeutet die Entdeckung von abgrenzbaren Gruppen von Daten, die möglichst viele Gemeinsamkeiten besitzen. Das Verfahren der Clusterbildung(Clustering, Clusteranalyse) bedeutet die (Teil-) Mengen/Haufen/Gruppen/Clustern zu identifizieren und die Daten entsprechend ihrer Zugehörigkeit zuzuordnen.

Ein Cluster oder eine Klasse ist dabei die Zusammenfassung von *ähnlichen* Objekten größtmöglicher Ähnlichkeit.

Cluster-Algorithmen haben das Ziel diese zusammengehörigen Klassen, welche möglichst viele Gemeinsamkeiten besitzen, automatisch zu bilden. Andersherum ausgedrückt erfolgt die Zuordnung zu einem Cluster so, dass verschiedene Cluster eine größtmögliche Unähnlichkeit besitzen.

Für Ähnlichkeit oder Unähnlichkeit gibt es definierte Maßzahlen, z.B. Distanzen

Beachte: Nach einer *Ballungsanalyse* sollte immer eine kanonische Benennung der Klassen erfolgen, das bedeutet die Klassenzuweisung muss erklärbar sein.

3.1. Clusterverfahren

Bei einer Klassifizierung interessiert der echte Wert der Variablen nicht, deswegen suche nichtlineare Transformation welche Klassifizierung erlaubt. Die Vorausetzung ist ein sinnvolles Distanzmaß. Problem: Bestimung der Anzahl der Cluster
Charakterisierungen von Clusterverfahren:

1. - agglomerativ: bottom up
   - divisiv: top down
2. - deterministisch => immer gleiche Klassen
   - stochastisch => verschiedene Klassen je Versuch
3. - inkrementell
   - nicht inkrementell → Erweiterung des Datensatzes möglich
4. - exhaustive: erschöpfende
   - nicht exhaustive: Möglich Datensatz keinem Cluster zuzuordnen
5. Klassenanzahl vorgegeben oder nicht gegeben
6. **Dichte- oder Abstandsbasierend**

Achtung: In jedem Clusterverfahren steckt eine Hypothese, wie die Cluster aussehen
Beachte: Untermannigfaltigkeit (n=1) nicht genau dann wenn Clustereverfahren (n>1), Projektionsverfahren passen nicht unbedingt auf Clusterverfahren.
Tipp: Strukturidee aus Projektionsverfahren => Clusterverfahren
Welche Typen von Klassenverfahren gibt es? Nenne einige Clusteralgorithmen.

3.1.1. hierarchisch: B/ ward

- n<4000 berechenbar (Grenze zur Anwendung)
- dargestellt durch *Dendrogramm*
- bottom up mit einem 1 Datenpunkt=1Cluster → Distanz zwischen zwei Clustern

Bem.: Jedes Dendrogramm ist eine Darstellung des *ultrametrischen* Anteils. Eine *Ultrametrik* ist im zweidimensionalen ein gleichschenkliges Dreieck ($d(x,z) = d(y,z)$).
Hierarchische Verfahren[1] erzeugen immer gleiches Dendrogramm, weil Verschmelzungsverfahren immer deterministisch sind. B/

- Single Linkage (SL): geometrisches Modell einer Kette, minimaler Abstand zw. 2 Clustern
- Complete Linkage (CL): geometrisches Modell einer Hyper-Kugel, maximaler Abstand zw. 2 Clustern
- Centroid Linkage: geometrisches Modell einer Hyper-Kugel, Abstand zwischen den Mittelvektoren der Cluster
- Average Linkage und Median Linkage: geometrisches Modell einer Hyper-Kugel, mittlere Distanz der Datenpunkte zwischen den Clustern
- Ward Methode („Kovarianz“): geometrisches Modell eines Hyper-Eies,
  - Varianzkriterium ist die Fehlerquadratsumme ΔQS
  - Verschmelze jeweils Cluster bei denen die Varianz minimal wird

3.1.2. partitionierend: B/ k means

Diese Verfahren sind disjunkt. **Achtung**: k means nicht genormt, d.h. Unterschiedliche Algorithmen in Matlab, SAP, SAS, R, dadurch unterschiedliche Ergebnisse

- partitionierend → alle Datenpunkte müssen zugeordnet werden
- geometrisches Modell: Kugel→ Überdeckung → Trennflächen(lineare Schnittflächen) → Minimierung
- optimiert die quadratischen Abweichungen von einem Mittelwert → Zentroiden
- Zentroide sind geometrische Mittel eines zweidimensionalen Bereiches, wobei dieser jeweilige Bereich entweder zufällig in der Datenwolke gewählt wird oder heuristisch bestimmt ist. Der Bereich ähnelt einer Voronoi-Zelle
  - vorgegeben wird k, also die Anzahl Zentroiden, sowie ihre (manchmal zufällige) Position
  - Liapunov-Funktion bekannt: $E^2 = \sum_{c=1}^{k} \sum_{i=1}^{n} (d(x_i) - m_c)^2$
  - *Gradient-Abstiegsverfahren*: Abstand d(Zentroid,Datensatz) wird minimiert
  - Position der Zentroide wird iterativ, also pro zufällig ausgewählten Datensatz, angepasst.
- analog zu EM ist die Wahl der Zentroide beim Start kritisch, d.h. lokales Minimierungsverfahren
- weiterer Knackpunkt ist die Veränderung der Zentroide
- **Nachteil**: Auswahl Startwerte und Wahl von Zentroiden, Ausreißer anfällig
- viele Details können je nach Implementierung unterschiedlich sein
- entspricht teilweise der SOM bei bestimmten Bedingungen (wenige Units)
- Fazit: Implementierung wichtig

3.1.3. überlappend: B/ Fuzzy

Emergente Verfahren mit unscharfe Bedingung genannt „Fuzzy“ => Zugehörigkeitsgrad zu einem Cluster p $\in [0,1]$, d.h. ein Datensatz kann mehreren Clustern zugeordnet werden.
B/ Bayes Clusterverfahren

[1] s. Carlsson 09 review

3.1.4. Dichte basierend: U^*C

Klassenbildung über U^*-Matrix mit : U^*C Clusterverfahren. U-Matrix bildet die Distanzabschätzung über die neuner Nachbahrschaft jeweils eines Neurons.

P-Matrix

- EM-Dichteschätzer: Naiv-Bayes→EM→Annahme von Unabhängigkeit, also Produkte in R^n
- Bilde Hyperkugel eines manuell zu wählendes Radiuses um Neuron. Anzahl Daten im R^n um dieses Neuron herum ergibt die Höhe
- nur bei richtiger Wahl des Radiuses der P-Matrix passen U-Matrix und P-Matrix zusammen

U^*-Matrix

Verechnung von U-Matrix und P-Matrix folgendermaßen:

- Bei einem Neuron in Umgebung hoher Dichten und niedrigen Distanzen wird das Tal zu einem See herabgesenkt => nur Dichten relevant
- Bei niedrigen Dichten und hohen Distanz wird der Hügel zu einem Berg erhoben =>nur Distanzen relevant

U^*C Clusteralgorithmus

watershed Transformation: emergentes Verfahren bei welchem keine lokale Höheneigenschaft sondern globale Kante nötig[2]. Die *Ballungsanalyse* erfolgt über die Wasserscheide: Man lasse es regnen und schaue wie das Wasser fließt: → Gradientenabstiegsverfahren:

- in U-Matrix vom Clusterrand, also Berg weg, da Cluster im Tal liegt
- in P-Matrix zur Clustermitte, also Berg hin, da Cluster auf Bergspitze vermutet wird

Auch durch bloßes Hinschauen möglich[3], indem Cluster von Bergen abgegrenzt werden. Gute Darstellung der Ergebnisse: *Politische Karte*

[2] B/„Marburg bis Meer“
[3] Für Publikation Automat notwendig

3.2. Gütemaße für Clusterverfahren

Falls Vorklassifikation bekannt, lässt sich die Qualität des Clusterns direkt vergleichen (s. Gl. **??**). Ansonsten ist der kritische Parameter die Anzahl der Cluster: Analog zur Approximation über Intervallbildung gilt, je mehr Cluster, desto bessere Qualität wird angezeigt. Deswegen sind verschiedene Anzahlen an Cluster schwer zu vergleichen.

Auch sollte die Stabilität der *Ballungsanalyse* durch n-fache Wiederholung des Verfahrens getestet werden. Allgemein gibt es die zwei folgenden Maße:

1. Homogenität: innere Clusterdistanz, wobei Maximum=Durchmesser.
   Bei **k-means**: $h(C_i) = \frac{1}{i} \sum_x Inner(i)$
   B/ mittleres Distanzmaß, Durchmesser, kleinster abstand, Varianz, Square Errors, etc
2. Heterogenität (zwischen Cluster): Minimum=Abstand
   B/ min, max, mittel

$$Güte \approx \sum(Hom) * \frac{1}{\sum(Het)} \tag{3.1}$$

Es gibt viele verschiedene Kennzahlen, welche beliebig gewählt werden können, da immer ein geometrisches Modell von Clustern dahinter steckt: B/

- Rand- Kriterium: Vergleiche Clusterverfahren über Ähnlichkeit $\in 1, 0, -1$, falls zwei Punkte in 2 Verfahren im selben Cluster sind. Rand Maß $\in [0, 1]$.
- Darstellung der Güte auch über Silhouette graphisch möglich, wobei geometrisches Modell die Kugel ist.[4]

Fazit: Clusterverfahren sind explorativ. Es gibt kaum Validierungsmöglichkeiten. Sie sind dadurch nur Optimierungen und die Voraussetzung ist Wissen zu dem jeweiligen Datensatz.

[4] Idee: Silhoutte mit anderen geometrischen Modell

Kapitel 4

Maschinelles Lernen - Klassifikatoren

KL sind Programme, die **neue** Datensätze in **bestehende** Klassen einordnen können. Eine Klassifikation ist eine eindeutige Zuordnung, d.h. kein Datensatz kann in zwei Clustern vorhanden sein.
F: Unterschied zwischen Clusteralgorithmen und Classifikatoren?
A:Erstes baut Cluster, zweiteres benötigt Cluster/Diagnose und ordnet neue Datensätze zu.

4.0.1. Wissensrepräsentationen

Grundkonzept der künstlichen Intelligenz (KI): passe deine Sprache dem Problem an. Definition in der KI: Wissen ist symbolische Repräsentation (sprachlich) von Objekten, Fakten, Regeln für einen Interpreter mit Symbolverarbeitungskompetenz. F: Worüber stellt man wissen dar?

1. Ontologien - Begriffshierarchien
   - Beziehung zu anderen Begriffen werden formell nicht festgehalten.
   - DAG: gerichteter(directed) acylic Graph: Graph mit„is a“ oder „part of“, B/ Zelle
2. Bäume
3. Regeln

F: Welche Punkte sollten bei der Wissensdarstellung beachtet werden?

- Einfachheit
- Gliederung: Wichtigstes zuerst, keine Wiederholungen
- Prägnanz: B/ 1-8 Hauptpunkte pro PowerPoint Folie
- Anrgegung: B/ Humor

4.1. Arten und Typen von KL

Es gibt zwei Arten von Unterteilungen:

1. symbolisch: Erzeugen Wissen
2. subsymbolisch: Können Klassifizieren

1. überwachtes Lernen
2. unüberwachtes Lernen

B/ subsymbolische Klassifikatoren

- NNK-Klassifikatoren
- Probabilistische Klassifikatoren, B/ Maximum Likelihood, Bayes nach EM,
- Neuronale Netze, B/ ESOM

Ab hier wird nur noch überwachtes Lernen der symbolischen Art behandelt. Es gibt drei Typen von regelbasierenden Verfahren zur Wissenserzeugung beim maschinellen Lernen:

1. **Entscheidungsbäume**:
   - Knoten enthalten Merkmale.
   - Kanten sind die zum Merkmal passenden Bedingungen.
   - Blatt stellt eine Klassenentscheidung dar.
   - Doch welches Merkmal soll zuerst betrachtet werden?
     - Kriterien für Merkmalsindex: Häufigkeit(GINI), Entropie (ID3, C4.5), Chi-Quadrat-Anpassung (CHAID)
2. **Entscheidungsregeln**:
   - Regeln aus Bäumen: 1 Regel entspricht dem Pfad von Wurzel bis Blatt
   - Danach Vereinfachung mittels Logik
   - Sprache ist Prolog: If Prämisse then Conclusion
3. **Assozationsregeln**:
   - Vor.: Effiziente Algorithmen wegen riesiger Datenmengen nötig B/ divide&conquer
   - *support* - Verbundswahrscheinlichkeit $P = p_1 * p_2 * p_3$ usw.
   - *confidence* - bedingte Wahrscheinlichkeit
   - *lift* - Theorem von Bayes
   - Ziel: Maximierung der oberen drei Dinge B/ A-priori-Algorithmus
   - Kritik: Regeln sinnvoll? B/Kaufvorschläge bei Amazon, Regel: „Bier und Windeln werden zusammengekauft“
4. **SIG*Algorithmus**:
   - Auf Regelnbasierender symbolischer Klassifikator
   - Input: Daten+Klassifizierung
   - Output: ?
   - deterministisch, beachte: kein Widerspruch mit *Fuzzy*-Kalkül, da fest ausgewählte Anzahl an relevanten Bedingungen[1]
   - Ziel: Verständlichkeit für Menschen bei hinreichender Präzession
     a) Charakterisierungsregelen:
        - Eine pro Klasse
        - Reihenfolge: Wichtigste Bedingung zuerst
     b) Differenzierungsregeln zwischen Klassen mit *Fuzzy*-Kalkül

[1] nach Wichtigkeit sortierte Bedingungen

Symbolische Klassifikatoren:

- Cart-Classifikation - Classification and Regression Tree
  - Entscheidungskriterium/Merkmalsauswahl über GINI-Index[2] $1 - \sum_i^c p_i(b)^2$ entspricht einer Häufigkeitsverteilung für jeden Konten b, s. Gl. **??**
  - *Pruning* - Baumbeschneidung
    - -> Level der Crossvalidation
    - => optimaler Cart => Wissensgewinn
  - Maß für Effiziens des Verfahrens: GINI-Gewinn $GINI(b)$-$GiniSplit(b)$
  - Nachteil: Numerische Werte sind schlecht zu behandeln
- Decision Stumps: viele schlechte Entscheider(Level 1 Bäume) werden gewichtet summiert

$$p_i(b) = \frac{Anzahl\ der\ durch\ Knotenausgewählten\ B/\ der\ Klasse\ i}{Anzahl\ aller\ durch\ Knote\ b\ ausgewählten\ B/} \tag{4.1}$$

F: Nenne weitere Algorithmen von symbolischen Klassifikationen, worin unterscheiden Sie sich?
A: Kriterien der Merkmalsauswahl, B/

- ID3: Informationsgewinn: Shanning-Info[3] $I = \sum_{alle} p * log(p)$ statt GINI
  - Numerische Werte sind schlecht zu behandeln
- C4.5: InformationsgainRatio: $I_{splitt} = \sum_{Nachfolger} p * log(p)$
  => $I_{gainratio}(b) = \frac{I:gain(b)}{I_{spliit}(b)}$, ergo Entropie
  - Verbesserung: Baum mit vielen Nachfolgern ist benachteiligt
    Analogie: MDS → Sammons Mapping

Vor Anwendung muss der Datensätze nach einer bestimmten Art und Weise in drei Teile aufgeteilt werden:

1. **Lerndaten**: Parameter des Kassifikators werden bestimmt.
2. **Testdaten**: Abschätzung der Generalisierungsfähigkeit eines Klassifikators → Güte
3. **Verifizierungsdaten**: Nur ein einziges mal verwendbar! => absolutes, endgültiges Gütemaß

Aufteilung zwischen Lern- und Testdaten wird n-mal durchgeführt, n>10. Dabei wird der Klassifikator mit den Lerndaten trainiert und die Qualitätsmessung wird auf dem Testdatensatz durchgeführt. Durch Wiederholungen erzeugt man eine Verteilung der Güte, über welche ein Mittelwert und Standardabweichung der Güte definierbar werden[4]. Methoden zum Aufteilen:

- *Split-Sample mit Quote*: Häufig 80 % Lerndaten und 20 % Testdaten
- *Cross-Validation*: bei kleinen Datenmengen
- *Bootstrapping*: zwischen 50 und 2.000 Durchläufe nötig

[2] Diversität
[3] Entropie
[4] Gelegentlich auch als crossvalidation bezeichnet.

4.2. Qualitätsbewertungen

Bem.: Nach 2 Theoremen, existert weder ein bester Klassifikator für alle Probleme noch eine optimale Merkmalsrepräsentation (s. auch Abb. **??**). Aber für diese maschinellen Entscheider ist in der Praxis eine optimale Datenrepräsentation entscheidend. Einfache Klassifikatoren sind häufig besser als komplexe (s. Occams Razor). Über n-malige Aufteilung der Daten erhält man eine Verteilungsannahme der *Accuracy*, welche die Qualität der *Ballungsanalyse* widerspiegelt. Andere Maße für Klassifikationsleistung sind ROC[5] (Receiver Operating Characteristik) und AUC (Fläche ROC). Begriffe:

- positive +, negative -, false F, true T $\rightarrow$ vier Kombinationen, daraus:
- Sensitivität: zurecht eingeordnet $\frac{T^+}{T^+ + F^-}$
- Spezifität: zurecht nicht zugeordnet $\frac{T^-}{T^- + F^+}$
- *Accuracy* (Korrektheit): $1 - x\%$ der Daten wurden falsch zugeordnet, bzw. $\frac{T^+}{N_{gesamt}}$

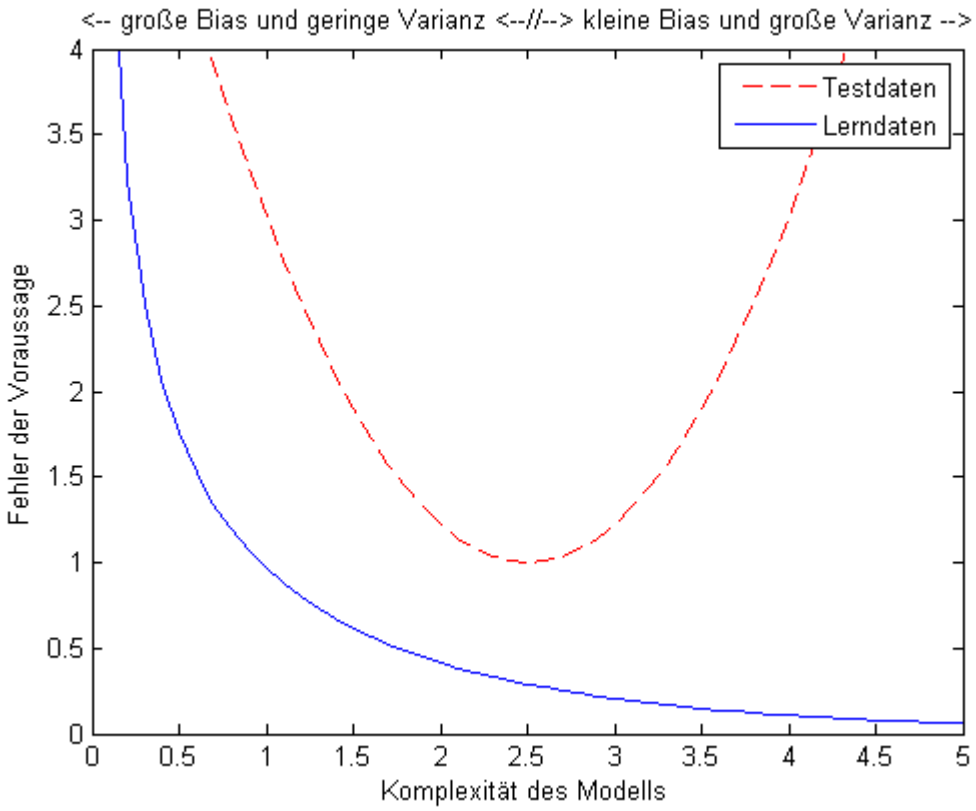


Abb. 4.2-1: Lernverhalten von Klassifikatoren beim Vergleich von Testdaten und Lerndaten. - Je mehr Epochen der Klassifikator lernt, desto komplexer wird das neuronale Netz und genauer die Anpassung der Units. Doch ab einer gewissen Grenze werden die zugrunde liegenden Prinzipien nicht gut abgebildet, also der Testdatensatz mit großen Fehlern eingeordnet.

[5]Sensitivität gegen Spezifität aufgetragen

Anhang

A. Glossar

Ähnlichkeitsmaß definiert entweder über nächster Nachbar Verfahren mit Parameter k der Anzahl oder über eine Distanz, Ziel ist Einordnung der Daten. 5

A-priori Anfangswahrscheinlichkeit aufgrund von Vorwissen, ergibt eine Dichtefunktion.. 3
Accuracy Übereinstimmmung des Baumes mit wahrer Klassenzuweisung. 18
Adjazenzmatrix Binäre Matrix, speichert welche Knoten des Graphen durch eine Kante verbunden sind. Eintrag in der i-ten Zeile und j-ten Spalte gibt an, ob eine Kante von dem i-ten zu dem j-ten Knoten führt. 9

Ballungsanalyse auch Clusteranalyse: Verfahren zur Entdeckung von Ähnlichkeitsstrukturen in (großen) Datenbeständen. Die gefundenen Gruppen von ähnlichen“ Objekten werden als Cluster bezeichnet, die Gruppenzuordnung als Clustering.. 12, 14, 18
Best Matching Unit Unit, das einem vorgegeben Wert am nächsten kommt. Alle benachbarten Units Lernen: Sie ändern ihren Gewichtsvektor.. 10, 12
Bootstrapping Bilden von Stichproben mit Zurücklegen: zufälliges Ziehen einer Menge von Daten als Lerndatenmenge. Die Daten verbleiben zusätzlich in der Menge der Testdaten.. 17

Cross-Validation k gleich große Teilmengen mit k-1 Lerndaten, und einem Testdatensatz.. 17

Databot Repräsentiert Datensatz und führt solange Schelling-Modell aus, bis nichtmehr gesprungen wird, dann wird Nachbarschaft um ϵ verkleinert. 10
Datensatz Eindeutig indizierte Anzahl an Werten verschiedener Merkmale. B/ Größe, Haarfarbe, Augenfarbe einer Person im Personalausweis. 1
Dendrogramm Baumstruktur: Die Wurzel repräsentiert die gesamte Menge, die Blätter des Baumes Klassen, in denen sich je ein einzelnes Objekt der Datenmenge befindet. Ein innerer Knoten ist die Vereinigung aller seiner Kindknoten. Jede Kante zwischen einem Knoten und einem seiner Kindknoten hat als Attribut die Distanz zwischen den beiden Mengen von Objekten. 13
Dichteverteilung Wahrscheinlichkeit der Variable einen Wert einzunehmen. 1

Emergenz spontane Herausbildung von neuen Eigenschaften oder Strukturen eines Systems infolge des Zusammenspiels seiner Elemente, wobei emergenten Eigenschaften des Systems nicht offensichtlich auf Eigenschaften der Elemente (El) zurückzuführen sind, die die El isoliert aufweisen. 10
Epoche Ein Durchlauf durch alle Daten. 10

Fuzzy Schätzkalkül analog zu Bayes oder Dempster. Im Gegensatz zu Bayes(Wahrscheinlichkeit/Zahl) ist hier aber die Datenstruktur eine Funktion. B/ für eine Regel müssen 5 von 7 Bedingungen erfüllt werden.. 16

geodätische Distanz zwischen KNN wird ein Weg von Punkt zu Punkt gezeichnet, also nicht direkte Distanz. B/ swiss roll. 5, 9
GMM Gaussmixturen Modell, anwendbar auf Verteilung mehrerer Moden. 2, 4
Gradient-Abstiegsverfahren Lösung für Optimierungsproblem der Numerik. Näherungswert wird in Richtung eines negativen Gradienten angepasst, bis optimaler Wert erreicht wird, also Verfahren konvergiert. 8, 13

KNN k nächster Nachbar. 9, 10
Kovarianzmatrix Jedes Element(i,j) der symmetrischen Matrix entspricht einer Produkt-Moment-Korrelation zwischen zwei Variablen (i,j). Die Diagonalelemente enthalten die Varianzen.. 3, 4
Kullback Leibler Divergenz Ähnlichkeitsmass zweier Funktionen, unsymmetrisches Integral, besitzt „crowding problem“. 11

Lernrate auch Abkühlung (cooling function), Anpassung der Gewichte der Units in einer Nachbarschaft. Geschieht entweder pro Datenwert nacheinander (Online SOM) oder Epochenweise (für alle Daten gleichzeitig). 10

Mode auch Gruppe, Cluster, oder Maximum bzw. Hügel einer Verteilung, Menge von Objekten mit ähnlichen Eigenschaften. 4

Politische Karte die Einfärbung der von den Datensätzen eingenommen Fläche entsprechend der Clusterung.. 12
Posteriori $P(G = g, X = x) = p(g|x))$ gibt den Wert an, dass der Datensatz x der Klasse g zugeordnet ist, Ergebnis des Theroems von Bayes, Wahrscheinlichkeitsverteilung. Dadurch wird der Wissensstand über einen unbekannten Umweltzustand nach der Beobachtung beschrieben.. 3
Potenzleiter Ladder of Powers nach [Tukey, 1977]. 4

Shepard Diagramm Scatterplot von Eingabe versus Ausgabe, Achtung ist Eingabe=Ausgabe, so ist das Plot optimal aber nicht aussagekräftig. 8

Split-Sample mit Quote Aufteilung durch zufälliges Ziehen eines bestimmten Prozentsatzes/Verhältnisses zwischen Lern- und Testdaten bei Einbezug der Klassenverteilung, also je Klasse. 17

U-Matrix Umgebungsmatrix, Höhendarstellung lokal integrierter Distanzen, also der Nachbarschaft. Aus großer Höhe folgt viel Platz im Eingaberaum und dadurch große Abstände. Täler bedeutet große Gemeinsamkeiten, Berge hohe Unterschiede.. 5

Ultrametrik Metrik mit schärfer Dreiecksungleichung: $d(x, y) \leq max(d(x, z), d(y, z))$. 5, 13

Units n-dimensionaler gewichteter Eigenvektor. 10

Variable Zuordnung von Werten zu einem zu untersuchenden Merkmal/Attribut jede Variable definiert eine Dimension. 1

Voronoi-Zelle Ist eine Zerlegung des Raumes in Regionen, welche alle Punkte des Raumes, die in Bezug zur euklidischen Metrik näher an dem Zentrum der Region liegen umfasst. Ein Spezialfall der Physik ist Wigner-Seitz-Zelle, welche nur einen Gitterpunkt enthält und zwar in ihrem Zentrum. Alle Orte im Inneren der Wigner-Seitz-Zelle liegen diesem Gitterpunkt näher als den benachbarten Gitterpunkten. Ihre Entsprechung im reziproken Gitter ist die erste Brillouin-Zone.. 9

Z-Transformation auch Standardisierung, Transformation einer Zufallsvariablen (ZV), so dass die resultierende ZV den Erwartungswert 0 und die Varianz 1 besitzt. Kritisch ist die Definition der Standardabweichung.. 4