

# **Maschinelles Lernen**

## **Zusammenfassung**

Thomas Mohr

# Contents

<b>1</b>	<b>Grundlagen</b>	<b>4</b>
1.1	(Un)-überwachtes Lernen . . . . .	4
1.2	Inkrementelles Lernen . . . . .	4
1.3	Aktives Lernen . . . . .	4
1.4	Data cleansing . . . . .	4
1.5	Datensatz . . . . .	4
<b>2</b>	<b>Deskriptive Statistik</b>	<b>6</b>
2.1	Beschreibung von Daten . . . . .	6
2.2	Mittelwert . . . . .	6
2.3	Median und Midrange . . . . .	6
2.4	Modus . . . . .	6
2.5	Varianz und Schiefe . . . . .	7
2.5.1	Moment $k$ -ter Ordnung . . . . .	7
2.6	Quantil . . . . .	8
2.6.1	Interquantile range (IQR) . . . . .	8
2.7	Korrelation zwischen Attributen . . . . .	8
2.7.1	Kovarianz für numerische Daten . . . . .	8
2.7.2	Korrelationskoeffizienten für numerische Daten . . . . .	9
2.7.3	Rangkorrelationskoeffizient . . . . .	9
2.7.4	$\chi^2$ -Test . . . . .	9
2.8	Visualisierung . . . . .	10
2.8.1	Boxplots . . . . .	10
2.8.2	Histogramme . . . . .	11
2.8.3	Quantil-Plots . . . . .	12
2.9	Distanzen . . . . .	13
2.9.1	Distanz auf numerischen Attributen . . . . .	13
2.9.2	Distanz auf ordinalen Attributen . . . . .	13
2.9.3	Distanz auf nominalen Attributen . . . . .	14
2.9.4	Distanz auf binären Attributen . . . . .	14
2.9.5	Distanz auf gemischten Typen . . . . .	14
2.10	Dimensionsreduktion und Einbettung in den Vektorraum . . . . .	15
2.10.1	Multidimensionale Skalierung . . . . .	15
2.11	Hauptkomponentenanalyse . . . . .	15
<b>3</b>	<b>Regression</b>	<b>16</b>
3.1	Bewertung und Fehler . . . . .	16
3.1.1	Fitten eines Polynoms . . . . .	17
3.2	Overfitting . . . . .	18
3.3	$k$ -fold Kreuzvalidierung . . . . .	19
3.4	Evalutation der Modelle . . . . .	19
3.4.1	Wilcoxon Test . . . . .	19

3.4.2	Bootstrapping . . . . .	20
<b>4</b>	<b>Klassifikation</b>	<b>20</b>
4.1	Fehler bei Klassifizierung . . . . .	20
4.2	$k$ -NN Klassifizierer . . . . .	21
4.2.1	Verdichtungstechniken . . . . .	21
4.3	Fehler bei binärer Klassifizierung . . . . .	21
4.4	Entscheidungsbäume . . . . .	22
<b>5</b>	<b>Probabilistische Verfahren</b>	<b>23</b>
5.1	Satz von Bayes . . . . .	24
5.1.1	Summenregel (Wahrscheinlichkeit) . . . . .	24
5.1.2	Produktregel (Wahrscheinlichkeit) . . . . .	25
5.1.3	Satz von Bayes – im univariaten Fall . . . . .	26
5.1.4	Schätzung der a-priori Wahrscheinlichkeiten . . . . .	26
5.1.5	Dichteschätzer . . . . .	26
5.1.6	Verteilungen . . . . .	27
5.1.7	Satz von Bayes – im multivariaten Fall . . . . .	29
5.2	Gradientenabstieg . . . . .	30
5.3	Exponentielle Verteilungen . . . . .	30
5.3.1	Bernoulli Verteilung . . . . .	31
5.3.2	Normalverteilung . . . . .	31
5.4	Ausreichende Statistik . . . . .	32
<b>6</b>	<b>Clustering</b>	<b>32</b>
<b>7</b>	<b>Warenkorbanalyse</b>	<b>32</b>
<b>8</b>	<b>Analyse von Graphdaten</b>	<b>32</b>

# 1 Grundlagen

## 1.1 (Un)-überwachtes Lernen

- Eine **überwachte** Lernaufgabe liegt vor, wenn wir Beispiele haben, die das zu lernende Attribut bereits tragen (Zielvariable).
  - **Regression** im Fall von kontinuierlichen Werten (z.B.  $\mathbb{R}$ ) - eigentlich numerisch
  - **Klassifikation** im Fall von diskreten Labeln (z.B. *TRUE*, *FALSE*; ausgezeichnet, durchschnittlich, schlecht) - eigentlich nominal oder ordinal
- Eine **unüberwachte** Lernaufgabe liegt vor, wenn es kein Attribut gibt, das wir lernen wollen und für das wir bereits Beispiele haben.
  - Clustering, also die Unterteilung der Daten in eine Menge von Gruppen
  - Finden von Ausreißern

## 1.2 Inkrementelles Lernen

- Anstatt das Modell stets von Null an zu lernen, wird das alte Modell mit neuen Beispielen erweitert.

## 1.3 Aktives Lernen

- Aktive Lernverfahren erzeugen die Beispiele selbst, d.h., sie sagen dem Benutzer, welches Tupel benötigt wird.

## 1.4 Data cleansing

- Fehlende Werte auffüllen
- Rauschen aus den Daten entfernen
- Daten glätten
- Ausreißer entfernen
- Identische Tupel identifizieren
- Daten komprimieren

## 1.5 Datensatz

- Ein Datensatz ist eine Tabelle.
- Eine Instanz (auch Objekt) ist eine Zeile in dieser Tabelle.
- Ein Attribut ist ein Feld, das ein Merkmal des Objekts repräsentiert. Mögliche Arten von Attributen sind:

- nominal (kategorisch)
  - \* Keine sinnvolle Ordnung
  - \* Wir können nicht rechnen (z.B. Mittelwert, Median, Abstände).
- ordinal (sortierte Kategorien)
  - \* Sinnvolle Ordnung
  - \* Der Unterschied zwischen zwei Ausprägungen ist i.d.R. unbekannt.
- binär
  - \* Können nur zwei Werte annehmen
- numerisch
  - \* Messbare Quantitäten
  - \* Abstand zwischen zwei Werten kann quantifiziert werden.
  - \* Auf den Attributen kann gerechnet werden.
  - \* Wir unterscheiden:
    - diskrete Attribute (endliche oder abzählbar unendliche Menge von womöglichen Ausprägungen)
    - kontinuierliche Werte, reelle Zahlen
    - Attribute mit echtem Nullpunkt (Gewicht, Größe)
    - Attribute ohne echten Nullpunkt (Jahresangaben, Temperatur in °C)
- Ein Datensatz besitzt  $N$  Instanzen und  $d$  Attribute.
  - $x_i$  beschreibt die  $i$ -te Instanz.
  - $x_{ij}$  beschreibt das  $j$ -te Attribut der  $i$ -ten Instanz.
  - $x$  beschreibt einen  $d$ -dimensionalen Vektor.
  - Liegt eine überwachte Lernaufgabe vor, so ist das Label der  $i$ -ten Instanz  $t_i$ .

## 2 Deskriptive Statistik

### 2.1 Beschreibung von Daten

- Wir betrachten nun Spalten des Datensatzes, also z.B. Spalte  $j$ :

$$X_j = (x_{1j}, \dots, x_{Nj})$$

### 2.2 Mittelwert

- Der Erwartungswert (Mittelwert) macht Aussagen zur Lage (dem "Zentrum") der Daten:

$$\mu_j := \sum_{i=1}^N x_{ij} \cdot p(x_{ij}) = \frac{1}{N} \sum_{i=1}^N x_{ij}$$

- Ist eine Gewichtung vorhanden, so kann der gewichtete Mittelwert herangezogen werden:

$$\mu'_j := \frac{\sum_{i=1}^N w_i x_{ij}}{\sum_{i=1}^N w_i}$$

- Problematisch bei Ausreißern

### 2.3 Median und Midrange

- Der Median ist der mittlere Wert in der sortierten Folge  $X_j$ .
- Das mittlere Element muss nicht existieren.
  - Per Definition wählen wir dann als Median den Wert

$$\frac{1}{2}(x_{\frac{N}{2},j} + x_{\frac{N}{2}+1,j})$$

im Fall numerischer Daten.

- Im Fall von ordinalen Daten kann der Median das linke oder das rechte Element sein, oder jede mögliche Ausprägung dazwischen.
- Der Midrange ist das arithmetische Mittel von Maximum und Minimum von  $X_j$ .

### 2.4 Modus

- Der Modus ist die am häufigsten vorkommende Ausprägung. Somit ist der Modus auch für nominale Attribute berechenbar.
- Wird die maximale Häufigkeit für mehr als einen Wert angenommen, so gibt es mehr als einen Modus.
- Kommt jede Ausprägung maximal einmal vor, so ist der Modus nicht existent.

## 2.5 Varianz und Schiefe

- Über einen Vergleich von Modus, Median und Mittelwert können wir (erste) Aussagen zur Schiefe machen.
- Über Maximum und Minimum können wir die Ausbreitung bestimmen.
- Mit dem Moment 2-ter und 3-ter Ordnung können wir beides auch quantisieren.
- Das Moment  $k$ -ter Ordnung des  $j$ -ten Attributs ist definiert als:

$$m_j^{(k)} = E((X_j - \mu_j)^k)$$

mit  $E(X) = \sum_{1 \leq i \leq N} x_{ij} p(x_{ij})$  und  $\mu_j$  ist Erwartungswert von  $X_j$

### 2.5.1 Moment $k$ -ter Ordnung

- $k = 1$  :?
- $k = 2$  :  $Var(X_j) := E((X_j - \mu_j)^2) = E(X_j^2) - \mu_j^2$ 
  - Die Varianz gibt die erwartete quadratische Abweichung vom Mittelwert an.
  - Sie ist also ein Maß für die Streuung der Daten (um den Mittelwert).
  - Die Quadratwurzel der Varianz wird als Standardabweichung bezeichnet und mit  $\sigma$  symbolisiert.

- $k = 3$  :  $v(X_j) := E((X_j - \mu_j)^3)$ 
  - Die Schiefe ist eine Kennzahl für die Asymmetrie einer Verteilung:

$$v(X) = \frac{3(\bar{X} - \tilde{X})}{s}$$

- $v(X_j) < 0$ : Verteilung ist linksschief
  - $v(X_j) > 0$ : Verteilung ist rechtsschief
  - $v(X_j) = 0$ : Verteilung symmetrisch
- $k = 4$  :  $w(X_j) := E((X_j - \mu_j)^4)$ 
  - Die Kurtosis ist eine Kennzahl für die Wölbung einer Verteilung:

$$w(X) = \frac{1}{N} \sum_{i=1}^N \left( \frac{x_i - \bar{X}}{s} \right)^4$$

- $w(X) < 0$ : Verteilung ist platykurtisch (flachgipflig)
  - $w(X) > 0$ : Verteilung ist leptokurtisch (steilgipflig)
  - $w(X) = 0$ : Verteilung ist mesokurtisch (normalgipflig)

## 2.6 Quantil

- Zur Berechnung der Quantile wird  $X_j$  zunächst aufsteigend sortiert.
- Das  $k$ -te Quantil ist der Wert  $x$  aus  $X_j$ , so dass maximal  $\frac{k}{q}$  der Werte in  $X_j$  kleiner als  $x$  sind, und  $\frac{(q-k)}{q}$  größer; für  $0 < k < q$ .
- Es gibt somit  $(q - 1)$   $q$ -Quantile.
- Sei  $p := \frac{k}{q}$ . Dann ist das  $k$ -te  $q$ -Quantil von  $X_j$  definiert als:

$$x_{pj} := \begin{cases} \frac{1}{2}(x_{Np} + x_{Np+1}) & Np \text{ gerade} \\ x_{\lfloor Np+1 \rfloor} & Np \text{ ungerade} \end{cases}$$

### 2.6.1 Interquantile range (IQR)

- Ist definiert als  $IQR = Q3 - Q1$
- Es gibt an, wie die 50% der mittleren Daten streuen
- Der IQR kann zudem benutzt werden, um Ausreißer zu erkennen.
  - Berechne  $\Delta = 1.5 \cdot IQR$
  - Ein Ausreißer ist ein Wert, der
    - \* kleiner  $Q1 - \Delta$  ist.
    - \* größer  $Q3 + \Delta$  ist.
- $Q1, Q2, Q3, IQR$  sowie Minimum und Maximum können graphisch im Boxplot zusammengefasst werden.

## 2.7 Korrelation zwischen Attributen

- Wir betrachten nun einen (möglichen) Zusammenhang der Spalten  $X_i$  und  $X_j$ .
- Je nach Attribut existieren unterschiedliche Maße:
  - Korrelationskoeffizienten und Varianz für numerische Daten
  - Rangkorrelationskoeffizienten für ordinale Daten
  - $\chi^2$ -Test für nominale Attribute

### 2.7.1 Kovarianz für numerische Daten

- Erlaubt zu messen, wie stark sich zwei Variablen gemeinsam ändern
- Wir benötigen den Begriff des Erwartungswerts, der hier aber dem Mittelwert entspricht:

$$E(X_j) = \overline{X_j} = \frac{1}{N} \sum_{i=1}^N x_{ij}$$



- $Cov(X_i, X_j) = E((X_i - \overline{X_i})(X_j - \overline{X_j})) = E(X_i X_j) - \overline{X_i} \cdot \overline{X_j}$
- Tendieren  $X_i$  und  $X_j$  dazu sich gemeinsam zu ändern, so ist  $Cov(X_i, X_j)$  positiv, bei entgegengesetzter Änderung negativ.
- Das Maß ist nicht normalisiert.

### 2.7.2 Korrelationskoeffizienten für numerische Daten

- Der Korrelationskoeffizient ist normalisiert im Intervall  $[-1, 1]$ :

$$cor(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}$$

- Wir haben keine Korrelation bei einem Wert von 0.
- Positive (negative) Korrelation liegt bei positiven (negativen) Werten vor.

### 2.7.3 Rangkorrelationskoeffizient

- Der (Spearman) Rangkorrelationskoeffizient basiert auf den Rängen der Elemente; wir betrachten die Spalten  $X_i$  und  $X_j$ . Er wird berechnet als:

$$r_s(X_i, X_j) = \frac{\sum_{1 \leq k \leq N} (rank(x_{ki}) - \mu Rank(X_i)) \cdot (rank(x_{kj}) - \mu Rank(X_j))}{\sqrt{\sum_{1 \leq k \leq N} (rank(x_{ki}) - \mu Rank(X_i))^2} \sqrt{\sum_{1 \leq k \leq N} (rank(x_{kj}) - \mu Rank(X_j))^2}}$$

mit  $\mu Rank(X_i)$  ist der mittlere Rang in Spalte  $i$

- Der Rang wird aufsteigend anhand der Werte bestimmt. Der kleinste Wert nimmt dabei Rang 1 ein, der zweitkleinste Rang 2, usw. Tritt ein Wert mehrfach auf, so ergibt sich der Rang aus dem Arithmetischen Mittel.
- $r_s$  ist normalisiert in  $[-1, 1]$ .

### 2.7.4 $\chi^2$ -Test

- Seien  $a_1, \dots, a_c$  die  $c$  Werte, die das Attribut  $X_k$  aufweist,  $b_1, \dots, b_r$  die  $r$  Werte, die wir in der Spalte  $X_l$  finden.
- Berechne in  $o_{ij}$  die beobachtete Anzahl der Ereignisse, dass  $X_k$  den Wert  $a_i$  und  $X_l$  den Wert  $b_j$  gemeinsam annehmen.
- Wir können auch die erwartete Anzahl berechnen (für nicht korrelierte Attribute):

$$e_{ij} = \frac{1}{N} (|X_k = a_i| \cdot |X_l = b_j|)$$

- Die Pearson  $\chi^2$  Statistik kann wie folgt berechnet werden:

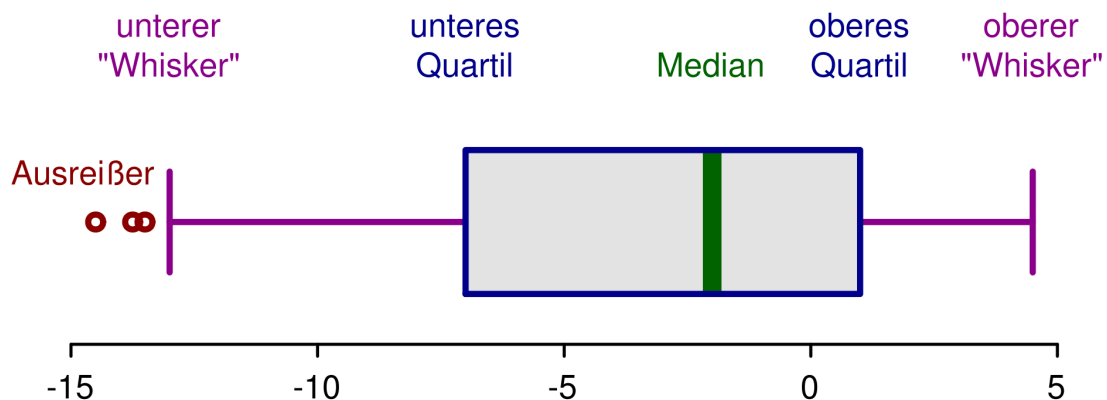
$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Die Statistik testet die Null-Hypothese der Unabhängigkeit zweier Variablen.
- Der Test basiert auf einem Signifikanzniveau mit  $(r - 1) \cdot (c - 1)$  Freiheitsgraden.
  - Das Signifikanzniveau ist die Wahrscheinlichkeit, mit der die Nullhypothese fälschlicherweise verworfen wird kann, obwohl sie eigentlich richtig ist.
- Die Hypothese kann abgelehnt werden, wenn der Wert der Prüfgröße größer ist als das  $(1 - \alpha)$ -Quantil der  $\chi^2$  Verteilung.

## 2.8 Visualisierung

### 2.8.1 Boxplots

- *IQR* ist die breite Mitte der Box.
- Das untere Quartil ( $X_{0,25}$ ) ist die untere/linke Kante der Box.
- Das obere Quartil ( $X_{0,75}$ ) ist die obere/rechte Kante der Box.
- Der Median ist durch eine Linie in der Box gekennzeichnet.
- Die langen Enden der Box heißen Whisker und geben die Grenzen für Ausreißer an. Alle Werte die außerhalb der Whisker, und damit des zulässigen Bereichs liegen, heißen Ausreißer.

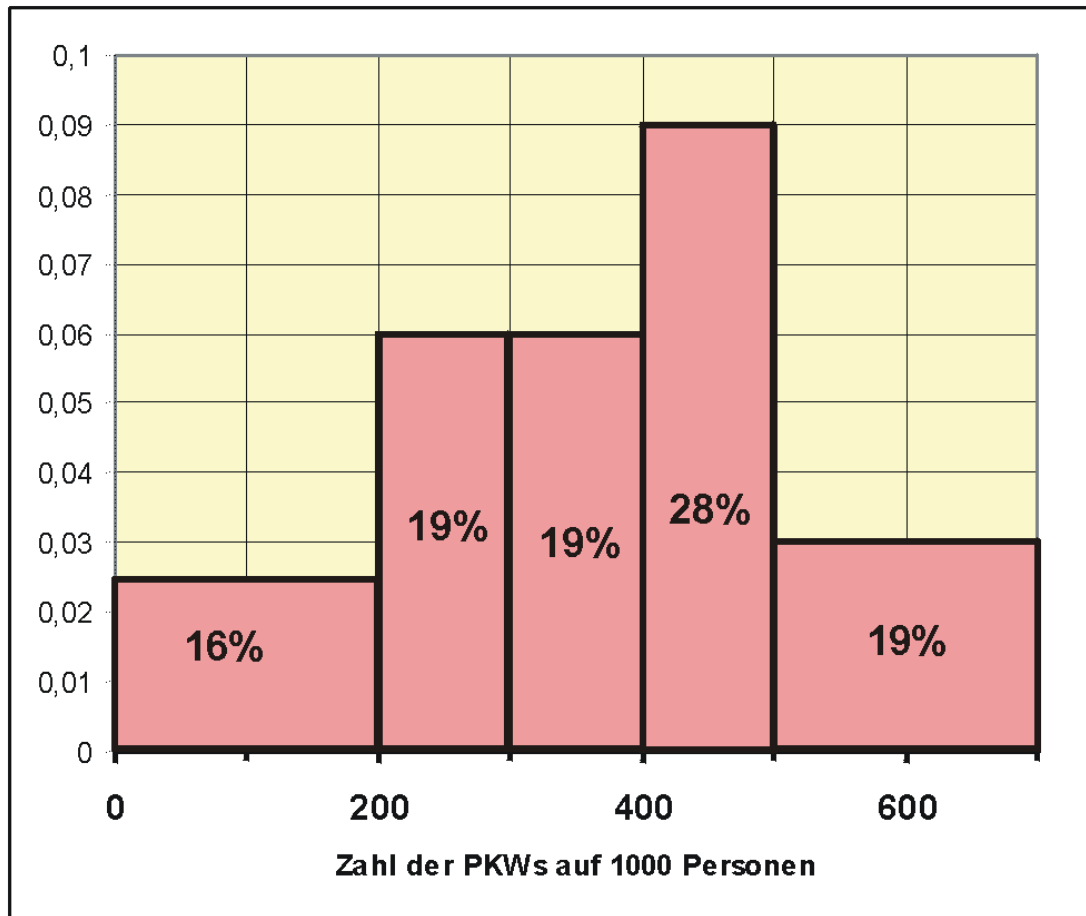


### 2.8.2 Histogramme

- Werden zur Darstellung von Häufigkeitsverteilungen verwendet.
- Bei numerischen Attributen müssen disjunkte Klassen definiert werden.
  - Die Balkenbreite kann durch zwei Verfahren bestimmt werden:
    - \* Scott-Regel:  $w = \frac{3,49 \cdot \sigma}{\sqrt[3]{N}}$
    - \* Regel von Diaconis:  $w = \frac{2(Q3-Q1)}{\sqrt[3]{N}}$
  - Die Häufigkeit ist proportional zum Flächeninhalt.

#### Beispiel

Klasse $j$	Zahl der PKW pro 1000	Anzahl der Länder (absolute Häufigkeit) $n_j$	Klassen- breite $d_j$	Rechteckshöhe (Häufigkeitsdichte) $h_j = \frac{n_j}{d_j}$
1	0 – 200	5	200	0,025
2	200 – 300	6	100	0,06
3	300 – 400	6	100	0,06
4	400 – 500	9	100	0,09
5	500 – 700	6	200	0,03
Summe $\sum$		32		



### 2.8.3 Quantil-Plots

- Ein Quantil-Plot erlaubt es das Verhalten der Werte eines Attributs abzuschätzen.
- Die Daten im  $i$ -ten Attribut werden sortiert und das  $k$ -te Element wird abgetragen auf  $f_k = \frac{k-0,5}{N}$ .

### Quantil-Quantil-Plots (qq-Plots)

- Die Quantile einer Verteilung werden gegen die Quartile einer anderen Verteilung abgetragen.
- Die Werte in werden in den Attributen  $X_i$  und  $X_j$  sortiert.
- Enthalten beide Attribute die gleiche Anzahl an Elementen, so wird  $x_{ki}$  auf  $x_{kj}$  mit  $1 \leq k \leq N$  abgebildet.
- Ansonsten ist  $|X_i| < |X_j|$  und nur  $|X_i|$  Punkte können geplottet werden:

- $x_{ki}$  ist das  $\frac{k-0,5}{|X_i|}$  Quantil.
- Das  $\frac{k-0,5}{|X_i|}$  Quantil von  $X_j$  muss dann interpoliert werden.

## 2.9 Distanzen

- Ähnlichkeits- oder Distanzmaß, dass ein Objekt-Paar auf einen numerischen Wert abbildet
- Metrik:
  - Identität:  $d(x_i, x_j) = 0 \iff x_i = x_j$
  - Symmetrie:  $d(x_i, x_j) = d(x_j, x_i)$
  - Dreiecksungleichung:  $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$
  - $d(\cdot, \cdot)$  beschreibt hier ein Funktion und ist nicht mit der Anzahl an Attributen zu verwechseln.
- Eine Distanz kann in eine Ähnlichkeit und umgekehrt umgewandelt werden. Ist  $d : 0 \times 0 \rightarrow [0, 1]$ , so kann  $s(x_i, x_j) = 1 - d(x_i, x_j)$  definiert werden.

### 2.9.1 Distanz auf numerischen Attributen

- Minkowski Abstand (Metrik)

$$d_h(x_i, x_j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + \dots + |x_{id} - x_{jd}|^h}$$

- $h = 1$ : Manhattan Distanz
- $h = 2$ : Euklidische Distanz
- Supremum Distanz für  $h \rightarrow \infty$ , die zu  $\max_{1 \leq f \leq d} |x_{if} - x_{jf}|$  konvergiert

### 2.9.2 Distanz auf ordinalen Attributen

- Betrachtung der Ränge und einer darauf basierenden Abbildung.
- Sei  $M_f$  die Menge möglicher Ränge für das Attribut  $f$ .
- Ersetze Wert  $x_{if}$  durch dessen Rang  $r_{if} \in \{1, \dots, M_f\}$ .
- Nun kann mit den Rängen gearbeitet werden, allerdings sollte zuvor normalisiert werden:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1} \in [0, 1]$$

- Die  $z_{if}$  sind numerisch und können beispielsweise mit der Minkowski Distanz verglichen werden.

### 2.9.3 Distanz auf nominalen Attributen

- Werden Objekte durch  $d$  nominale Attribute beschrieben, so kann die Distanz zwischen  $x_i$  und  $x_j$  wie folgt berechnet werden

$$d(x_i, x_j) = \frac{d - m}{d}$$

wobei  $m$  die Anzahl der Übereinstimmungen ist.

### 2.9.4 Distanz auf binären Attributen

	1	0	$\sum$
1	$q$	$r$	$q + r$
0	$s$	$t$	$s + t$
$\sum$	$q + s$	$r + t$	$d$

- Je nachdem, ob Attribute symmetrisch sind, können zwei verschiedene Distanzen definiert werden.

- Ist sowohl der Zustand "0" als auch "1" gleichwertig, so definieren wir die Distanz als:

$$d(x_i, x_j) = \frac{r + s}{d}$$

- Im Fall eines asymmetrischen Attributs tragen die "1"-en die tatsächliche Information; "0"-en sind nicht von Interesse:

$$d(x_i, x_j) = \frac{r + s}{q + r + s}$$

- Der Jaccard-Koeffizient ist ein häufig vorkommendes Ähnlichkeitsmaß:

$$s(x_i, x_j) = 1 - d(x_i, x_j) = \frac{q}{q + r + s}$$

### 2.9.5 Distanz auf gemischten Typen

- Sei  $d$  die Anzahl unterschiedlicher Attributstypen:

$$d(x_i, x_j) = \frac{\sum_{f=1}^d \delta_{ij}^{(f)} \frac{|x_{if} - x_{jf}|}{\max_{1 \leq h \leq N} x_{hf} - \min_{1 \leq h \leq N} x_{hf}}}{\sum_{f=1}^d \delta_{ij}^{(f)}}$$

- $\delta_{if}^{(f)}$  ist ein binärer Indikator.
  - Er ist 0, falls ( $x_{if}$  oder  $x_{jf}$  unbekannt sind, oder wenn)  $x_{if} = x_{jf} = 0$  und das binäre Attribut  $f$  asymmetrisch ist; ansonsten ist  $\delta_{if}^{(f)} = 1$ .

## 2.10 Dimensionsreduktion und Einbettung in den Vektorraum

### 2.10.1 Multidimensionale Skalierung

- Überführung von Punkten aus einem  $d$ -dimensionalen Raum in einen  $m$ -dimensionalen Raum ( $d > m$ ) oder metrischer Raum in Vektorraum.
- Die paarweisen Euklidischen Abstände sollen dabei möglichst wenig verändert werden.
- Es gilt:

$$\begin{aligned} d(x_i, x_j)^2 &= d_{ij}^2 = \sum_{k=1}^d (x_{ik} - x_{jk})^2 \\ &= \underbrace{\sum_{k=1}^d (x_{ik})^2}_{b_{ii}} - 2 \underbrace{\sum_{k=1}^d x_{ik} x_{jk}}_{b_{ij}} + \underbrace{\sum_{k=1}^d (x_{jk})^2}_{b_{jj}} \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned}$$

- Zentriere Daten im Ursprung:  $\sum_{i=1}^N x_{ij} = 0 \quad \forall j = 1, \dots, d$
- Die  $b_{ij}$  können zu einer  $(N \times N)$ -Matrix  $B$  zusammengefasst werden. Daher gilt  $B = XX^T$ .
- $X$  ist die gesuchte Matrix, die den Datensatz durch  $N$  Attribut-Vektoren beschreibt und die es nun zu approximieren gilt.
- Spektrale Zerlegung:

$$X = CD^{\frac{1}{2}}$$

mit  $C$  ist die Matrix, deren Spalten den Eigenvektoren von  $B$  entsprechen;  $D$  ist eine diagonale Matrix mit den Eigenwerten.

### 2.11 Hauptkomponentenanalyse

- Die Hauptkomponentenanalyse (PCA) projiziert ein Objekt  $x \in \mathbb{R}^d$  auf  $z \in \mathbb{R}^d$  wie folgt:

$$z = w^T x$$

- Ziel ist es durch eine Projektion die Varianz auf den neuen Attributen  $Z_1, \dots, Z_d$  zu maximieren.
- Tatsächlich beträgt dabei die Korrelation zwischen allen Paaren  $(Z_i, Z_j)$  auch 0.
- Gesucht ist ein neuer  $m(< d)$  dimensionaler Raum, auf dem die Daten mit minimalem Informationsverlust projiziert werden können.

### 3 Regression

- Es liegen numerische Daten vor.
- Es existiert eine Zielvariable, die wir aus den anderen hervorgesagt werden soll.
- Ein Modell

$$f : \mathbb{R}^d \rightarrow \mathbb{R}$$

soll gelernt werden.

- Das Problem wird als
  - univariat bezeichnet, falls  $d = 1$ .
  - multivariat bezeichnet, falls  $d > 1$ .
- Ein Modell  $y(x)$  muss bewertet werden können.
- Dazu wird eine Fehlerfunktion  $L : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  benötigt, die den Fehler auf den zukünftigen Eingaben misst.
- Eine gängige Wahl für die Regression ist der quadratische Fehler:

$$(y(x), t) \rightarrow (y(x) - t)^2$$

- Das Risiko (der erwartete Fehler) kann somit wie folgt angegeben werden:

$$R(y) = E[L] = \int L(t, y(x)) dP(x, t)$$

- Dieses Risiko kann jedoch nicht berechnet werden.

#### 3.1 Bewertung und Fehler

- Die Approximation  $R(y) = \int L(t, y(x)) dP(x, t)$  führt zum empirischen Risiko:

$$\begin{aligned} R_{emp}(y) &= \frac{1}{N} \sum_{i=1}^N L(y(x_i), t_i) \\ &= \frac{1}{N} \sum_{i=1}^N (y(x_i) - t_i)^2 \end{aligned}$$

- Dieser Ausdruck kann ausgewertet werden. Es wird eine Funktion (ein Modell)  $y : \mathbb{R}^d \rightarrow \mathbb{R}$  gesucht, die das empirische Risiko minimiert.



### 3.1.1 Fitten eines Polynoms

- Nun wird der multivariate Fall betrachtet ( $x_{i0} = 1$  für  $1 \leq i \leq N$ ).
  - Somit wird eines neues Attribut  $X_0$  hinzugefügt, mit Wert 1 für jede Instanz.

$$\begin{pmatrix} X_0 & \dots & X_d \\ 1 & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_{Nd} \end{pmatrix}$$

$$\begin{aligned} y(x_i, w) &= w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + \dots + w_d x_{id} \\ &= \sum_{j=0}^d w_j x_{ij} \end{aligned}$$

- Nun muss das optimale  $w$  gefunden werden, also jenes, für das

$$R_{emp}(w) = \frac{1}{N} \sum_{i=1}^N L(y(x_i), t_i)$$

minimiert wird.

- Gesucht wird also  $w^* = A^{-1}y$  mit:

$$A = \begin{pmatrix} \sum_i x_{i0}x_{i0} & \sum_i x_{i1}x_{i0} & \sum_i x_{i2}x_{i0} & \dots & \sum_i x_{id}x_{i0} \\ \sum_i x_{i0}x_{i1} & \sum_i x_{i1}x_{i1} & \sum_i x_{i2}x_{i1} & \dots & \sum_i x_{id}x_{i1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{i0}x_{id} & \sum_i x_{i1}x_{id} & \sum_i x_{i2}x_{id} & \dots & \sum_i x_{id}x_{id} \end{pmatrix}$$

$$w = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$y = \begin{pmatrix} \sum_i t_i x_{i0} \\ \sum_i t_i x_{i1} \\ \sum_i t_i x_{i2} \\ \vdots \\ \sum_i t_i x_{id} \end{pmatrix}$$

- Die Berechnung kann effizienter gestaltet werden durch  $w^* = (D^T D)^{-1} D^T t$  mit:

$$D = \begin{pmatrix} x_{i0} & x_{11} & x_{12} & \dots & x_{1d} \\ x_{i0} & x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{i0} & x_{N1} & x_{N2} & \dots & x_{Nd} \end{pmatrix}$$

$$t = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_N \end{pmatrix}$$

### 3.2 Overfitting

- Es wird nicht nur das durch die Daten zugrundeliegende Modell, sondern auch das Rauschen, gelernt.
- Jedoch soll ein Modell erzeugt werden, das gut generalisiert.
- Es kann ein Regularisierungsterm verwendet werden, um hohe Koeffizienten zu bestrafen:

$$R'(w) = \frac{1}{2} \sum_{i=1}^N (y(x_i, w) - t_i)^2 + \frac{\lambda}{2} \|w\|^2$$

- Ein  $\lambda = 0$  führt zu dem alten Ansatz; je größer  $\lambda$ , desto stärker werden hohe Koeffizienten bestraft.
- Es soll eine gute Generalisierung erreicht werden, demzufolge muss

$$\int L(t, y(x)) dP(x, t)$$

minimiert werden.

- Es wird ein Datensatz zum Lernen und einer zum Validieren benötigt.
- Ist nur ein Datensatz gegeben, so kann die  $k$ -fold Kreuzvalidierung Anwendung finden.
- Eine weitere Methode (bei wenigen Daten) ist das Bootstrapping.

### 3.3 $k$ -fold Kreuzvalidierung

- Die Daten werden zufällig permutiert und in  $k$  (annähernd) große Buckets verteilt.
- Es wird beginnend bei  $i = 1$  das Bucket  $i$  beiseite gelegt.
- Die verbleibenden Buckets werden als Trainingsdaten verwendet; das beiseite gelegte als Testdatensatz.
- Somit ergeben sich  $k$  Ergebnisse, mit denen die Modelle bewertet werden können.
- Aggregation z.B. durch Mittelwert und Standardabweichung führt zu Punktschätzer.

### 3.4 Evaluation der Modelle

#### 3.4.1 Wilcoxon Test

- Es werden zwei Stichproben danach getestet, ob
  - der Mittelwert der einen Stichprobe kleiner-gleich dem Mittelwert der anderen Probe ist (einseitiger Test):

$$H_0 : \mu_1 \leq \mu_2$$

und

$$H_1 : \mu_1 > \mu_2$$

- die Mittelwerte identisch sind (zweiseitiger Test):

$$H_0 : \mu_1 = \mu_2$$

und

$$H_1 : \mu_1 \neq \mu_2$$

- Für den Test müssen folgende Stichprobenvariablen berechnet werden ( $R_{\cdot,1}$  ist der Vektor der empirischen Fehler der ersten Parametrisierung,  $R_{\cdot,2}$  analog):

$$D_i = R_{i,1} - R_{i,2}$$

- Berechnet werden folgende Werte:

$$rg_i = \text{rang}(|D_i|)$$

$$W_+ = \sum_{i=1}^N \mathbb{I}_{R_{i,1} - R_{i,2} > 0} rg_i$$

$$W_- = \sum_{i=1}^N \mathbb{I}_{R_{i,1} - R_{i,2} < 0} rg_i$$

$$W = \mathbb{I}_q = \begin{cases} 1 & q \\ 0 & \neg q \end{cases}$$

- Gilt  $R_{i,1} - R_{i,2} = 0$ , so wird das Paar keinem der Werte  $W_+$  und  $W_-$  zugeordnet.

## Beispiel

$R_1$	$R_2$	$D_i$	$ D_i $	$rg_i$	$W_+$	$W_-$
5	8	-3	3	2,5		2,5
3	10	-7	7	5		5
15	12	3	3	2,5	2,5	
25	20	5	5	4	4	
18	19	-1	1	1		1

- Berechne Minimum aus den Summen von  $W_+$  und  $W_-$ :

$$\min\{6.5, 8.5\} = 6.5$$

- $\forall i. R_{i,1} - R_{i,2} \neq 0 \implies n = N = 5$

### 3.4.2 Bootstrapping

- Bei sehr kleinen Datensätzen würden die Folds (Kreuzvalidierung) sehr klein werden.
- Daher werden zufällig  $N$  gleichverteilte Instanzen aus dem Datensatz der Größe  $N$  gezogen; das Ziehen erfolgt mit Zurücklegen.
- Diese Prozedur kann  $k$ -mal wiederholt werden, um mehrere Trainings- und Testdatensätze zu erzeugen.
- Es gilt:

$$\underbrace{\left(1 - \frac{1}{N}\right)^N}_{\substack{\text{Instanz } x_i \text{ wird nach} \\ N \text{ Ziehungen nicht gezogen}}} \approx e^{-1} = 0,368$$

- Somit enthält der Trainingsdatensatz 63,2% der Instanzen.

## 4 Klassifikation

### 4.1 Fehler bei Klassifizierung

- Auf analoge Weise zur Regression ergibt sich das Risiko für die Klassifikation:

$$\sum_k \sum_j \int_{R_j} L_{kj} P(x, k) dx$$

–  $k, j$  sind Klassen

- $R_j$  sind Klassenregionen
- Der Fehler kann asymmetrisch sein
- Erneut kann das Risiko nicht ausgewertet werden und daher wird der empirische Fehler (mit  $\hat{t}_i$  ist prognostizierte Klasse) bestimmt:

$$\frac{1}{N} \sum_{i=1}^N L_{\hat{t}_i, t_i} \quad \overset{\text{symmetrische Variante}}{\approx} \quad \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{t_i \neq \hat{t}_i}$$

## 4.2 $k$ -NN Klassifizierer

- Anstatt ein Modell zu lernen, werden im Beispiel der Trainingsdaten die  $k$  ähnlichsten Objekte gesucht, und damit  $k$  Ausprägungen des zu lernenden Attributs.
- Basierend auf den  $k$  Ausprägungen wird eine Mehrheitsentscheidung getroffen.
- Dazu wird ein geeignetes Ähnlichkeitsmaß benötigt, z.B. ein passendes  $h$  bei der Minkowski-Distanz.
- Bei großen Datensätzen kann der  $k$ -NN Klassifizierer ineffizient werden.

### 4.2.1 Verdichtungstechniken

- Bestimmte Instanzen definieren stückweise lineare Funktionen (Voronoi-Diagramm), die zur Klassifikation gebutzt werden können.
- Die Instanzen werden greedy bestimmt; Ausgang ist 1-NN.
  1. Initial ist die Menge  $Z$  der gesuchten Instanzen leer
  2. Durchlaufen der Instanzen  $x$  des Datensatzes in jedem neuen Zyklus in neuer zufälliger Reihenfolge (bis  $Z$  stabil ist) und betrachte  $x_i$ :
    - a) Finde das Element  $x_j$  in  $Z$ , das die minimale Distanz zu  $x_i$  ausweist (ist  $Z$  noch leer, so füge  $x_i$  zu  $Z$  hinzu).
    - b) Weist  $x_i$  nicht das selbe Label auf wie  $x_j$ , so füge  $x_i$  zu  $Z$  hinzu.

## 4.3 Fehler bei binärer Klassifizierung

- Andere Maße vergleichen
  - *wahr positiv* (tp)
  - *wahr negativ* (tn)
  - *falsch positiv* (fp)
  - *falsch negativ* (fn)

	Predicted class		
True class	Positive	Negative	Total
Positive	$tp$ : true positive	$fn$ : false negative	$p$
Negative	$fp$ : false positive	$tn$ : true negative	$n$
Total	$p'$	$n'$	$N$

- Die ROC-Kurve trägt für verschiedene Parametrisierungen eines Algorithmus (z.B. Loss-Matrix)  $\frac{fp}{n}$  gegen  $\frac{tp}{p}$  ab.
- Interessant ist vor allem die AUC, also die Fläche unter der ROC-Kurve. Ist diese eins, so liefert der Klassifizierer ein optimales Ergebnis.

#### 4.4 Entscheidungsbäume

- Der Entscheidungsbaum ist ein hierarchisches Modell.
- Es werden lokale Regionen durch eine Sequenz von Aufteilungen identifiziert.
- Jeder Knoten definiert eine Testfunktion mit einem diskreten Ergebnis.
- Es wird an der Wurzel gestartet und ein Durchlauf entlang eines Pfades bis zum Blatt wird gestartet.
- Die Baum-Induktion erfolgt durch eine Stichprobe (Trainingsdaten); die Verfahren sind greedy und suchen in jedem Schritt die lokal beste Aufteilung.
- Die Zahl der Elemente, die Knoten  $m$  erreichen, sei  $N_m$  (in der Wurzel ist diese Zahl  $N$ ).
- $N_m^{(i)}$  der Elemente gehört zu Klasse  $C_i$  und somit gilt:

$$\sum_i N_m^{(i)} = N_m$$

- Die Schätzung am Knoten  $m$  beträgt:

$$\begin{aligned}\hat{P}(C_i | x, m) &\equiv p_m^{(i)} \\ &= \frac{N_m^{(i)}}{N_m}\end{aligned}$$

- Ein Knoten ist rein, wenn die Schätzung entweder 0 oder 1 ist.
- Bei reinen Knoten ist keine weitere Zerlegung notwendig; es wird ein Blatt gebildet.
- Eine Möglichkeit zur Messung der Unreinheit ist die Entropie:

$$I_m = - \sum_{i=1}^k p_m^{(i)} \log_2 p_m^{(i)}$$

mit  $\lim_{n \rightarrow 0} n \log_2 n = 0$  und daher  $0 \log_2 0 \stackrel{def}{=} 0$ .

- Eine uniforme Verteilung hat eine höhere Entropie als eine nicht-uniforme Verteilung.

### Beispiel

- Betrachtet wird der Knoten  $m$  (zu Beginn die Wurzel).
  - Welches Attribut soll zur nächsten Verzweigung gewählt werden?
  - Es werden univariate Bäume verwendet; multivariate bringen im Allgemeinen keine Vorteile.
- Angenommen es wird das Attribut  $a$ ,  $1 \leq a \leq d$  betrachtet.
  - Ist es numerisch, gibt es zwei Verzweigungen gemäß Test  $x_{ia} \leq \theta_0$ .
  - Ist es diskret, so gibt es so viele Verzweigungen, wie das Attribut (verschiedene) Ausprägungen hat.
  - Es gibt im Allgemeinen  $v$  Verzweigungen.
- Von den  $N_m$  Elementen, die Knoten  $m$  erreichen, nehmen  $N_{mj}$  die  $j$ -te der  $v$  Verzweigungen,  $N_{mj}^{(i)}$  davon gehören zur Klasse  $i$ .
- Für die Kinder von  $m$  können die Wahrscheinlichkeiten für die Klasse  $i$  ermittelt werden, für Kind  $j$  gilt insbesondere:

$$\begin{aligned}\hat{P}(C_i | x, m, j) &\equiv p_{mj}^{(i)} \\ &= \frac{N_{mj}^{(i)}}{N_{mj}}\end{aligned}$$

- Würde im Attribut  $a$  verzweigt werden, so ergibt sich bei einem  $k$ -Klassen Problem eine neue Entropie von:

$$I'_m = - \sum_{j=1}^v \frac{N_{mj}}{N_m} \cdot \sum_{i=1}^k p_{mj}^{(i)} \log_2 p_{mj}^{(i)}$$

## 5 Probabilistische Verfahren

- Es werden nun Wahrscheinlichkeitsverteilungen auf den Attributen und der Zielvariable betrachtet.
- Gegeben ist ein univariater Datensatz, anhand dessen Kunden, basierend auf dem Attribut *Einkommen*  $X_1$ , auf Kreditwürdigkeit hin klassifiziert werden sollen.
- Die Kreditwürdigkeit kann durch eine Bernoulli Variable dargestellt werden, bedingt durch die Variable  $X_1$ .
- $C = 1$  entspricht dabei hohem Ausfallrisiko, und  $C = 0$  einem geringem Ausfallrisiko

- Es könnte sogar die Fehlerwahrscheinlichkeit

berechnet oder eine Ablehnungsoption verwendet werden.

## 5.1 Satz von Bayes

- Mit dem Satz von Bayes kann  $P(C | x)$  berechnet werden:

$$P(C \mid x) = \frac{P(C)p(x \mid C)}{p(x)}$$

- $P(C)$  ist die **a-priori** Wahrscheinlichkeit.
- $p(x | C)$ , der **Klassen-Likelihood**, ist die Wahrscheinlichkeit, dass ein zu  $C$  gehörendes Ereignis den Beobachtungswert  $x$  hat.
- $p(x)$  ist die **Evidenz**, die Randwahrscheinlichkeit, dass die Beobachtung  $x$  gemacht wird (nicht direkt berechenbar).

### 5.1.1 Summenregel (Wahrscheinlichkeit)

A diagram of a 2D grid representing a matrix. The grid is 6 rows high and 9 columns wide. The horizontal axis is labeled  $x_i$  at the bottom center, with a bracket above it spanning the first five columns. The vertical axis is labeled  $y_j$  on the left side, with a bracket to its right spanning the first five rows. A specific cell in the third row and sixth column is labeled  $n_{ij}$ . A bracket above this cell spans the sixth column, labeled  $c_i$ . A bracket to the right of the third row spans the entire row, labeled  $r_j$ .

- Angenommen  $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$  ist bekannt.
- Es gilt

$$p(X = x_i) = \frac{c_i}{N} = \sum_j \frac{n_{ij}}{N}$$



und damit

$$\begin{aligned} p(X = x_i) &= \sum_j \frac{n_{ij}}{N} \\ &= \sum_j p(X = x_i, Y = y_j) \end{aligned}$$

### 5.1.2 Produktregel (Wahrscheinlichkeit)

$y_j$					$n_{ij}$				

$x_i$

$c_i$

$r_j$

- Tatsächlich ist bekannt:  $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$ .
- Ferner ist bekannt:  $p(X = x_i) = \frac{c_i}{N}$ .
- Daraus ergibt sich

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} \\ &= \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \end{aligned}$$

und somit

$$p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i)p(X = x_i)$$

- $P(C)$  und  $p(x | C)$  können basierend auf den Daten oder einer Stichprobe davon berechnet werden.
- $p(x) = p(x | C = 1)P(C = 1) + p(x | C = 0)P(C = 0)$  im Fall eines binären Klassifikationsproblems
- Im allgemeinen Fall gibt es  $k$  Klassen:

$$P(C_i | x) = \frac{P(C_i)p(x | C_i)}{\sum_k P(C_k)p(x | C_k)}$$

- Klasse  $k$  wird gewählt, falls  $k = \arg \max_i P(C_i | x)$

### 5.1.3 Satz von Bayes – im univariaten Fall

- Die Dichten der Verteilungen  $P(C_i)$  und  $p(x \mid C_i)$  müssen für alle  $i$  geschätzt werden.
- Es kann eine (bis auf die Parameter) bekannte Verteilung vorliegen.
  - Es gibt Tests, um auf eine bestimmte Verteilung hin zu testen.
  - Es reichen jedoch auch Histogramme und qq-Plots.
  - Die Berechnung der unbekannten Parameter erfolgt durch Optimierung (Maximum Likelihood).
- Die Verteilung kann sich aus mehreren bekannten Dichten zusammensetzen (z.B. Mixture of Gaussians).
- Wenn die Dichte nicht bekannt ist, so kann auf ein  $k$ -NN oder Kernel Verfahren zurückgegriffen werden.

### 5.1.4 Schätzung der a-priori Wahrscheinlichkeiten

- Die a-priori Wahrscheinlichkeiten werden aus dem Datensatz geschätzt mit

$$p(C_k) = \frac{N_k}{N}$$

wobei  $N_k$  die Anzahl der Instanzen mit der Klassenzugehörigkeit  $k$  ist und  $N$  die Anzahl der Daten im Datensatz.

### 5.1.5 Dichteschätzer

- Gegeben sind unabhängige und identisch verteilte Stichproben:

$$X = \{x_{i1}\}_{i=1}^N = \{x_i\}_{i=1}^N$$

- Die  $x_i$  sind nach einer bis auf  $\theta$  bekannten Dichte  $p(x \mid \theta)$  gezogen worden.
- Gefunden werden soll das  $\theta$ , bei dem die  $x_i$  am wahrscheinlichsten aus  $p(x \mid \theta)$  gezogen wurden.
- Aufgrund der iid Annahme ergibt sich die Likelihood:

$$\begin{aligned} l(\theta \mid X) &\equiv p(X \mid \theta) \\ &= \sum_{i=1}^N p(x_i \mid \theta) \end{aligned}$$

- Ziehen des Logarithmus (log-Likelihood) und Ableiten ermöglicht nun das Maximieren.

## Gauss-Verteilung

- $p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \frac{-(x-\mu)^2}{2\sigma^2}$
- Die Maximum Likelihood (ML) Schätzer sind

$$\hat{m} = \frac{\sum_{i=1}^N x_i}{N}$$
$$\hat{s}^2 = \frac{\sum_{i=1}^N (x_i - \hat{m})^2}{N}$$

## Bernoulli-Verteilung und deren Verallgemeinerung

- $P(x \mid p) = p^x(1-p)^{1-x}, x \in \{0, 1\}$
- Der ML Schätzer ist  $\hat{p} = \frac{\sum_{i=1}^N x_i}{N}$
- Verallgemeinert auf  $k$  Zustände erhält man ( $\sum_{j=1}^k p_j = 1$ )

$$P(x_{i1}, \dots, x_{ik} \mid p) = \prod_{j=1}^k p_j^{x_{ij}}$$

und als ML Schätzer

$$\hat{p} = \frac{\sum_{i=1}^N x_{ij}}{N}$$

## Binomialverteilung

- Verwandt mit dem Bernoulli Experiment
- $m$  ist die Anzahl der Beobachtungen mit  $x = 1$  für ein Bernoulli Experiment (bzw. die zugehörige Variable)
- $\text{Bin}(m \mid N, p) = \binom{N}{m} p^m (1-p)^{N-m}$
- $E(m) = Np$
- $\text{Var}(m) = Np(1-p)$

### 5.1.6 Verteilungen

- Setzt sich eine Verteilung aus  $n$  Verteilungen (z.B. Normalverteilungen) zusammen, so gilt:

$$p(x) = \sum_{j=1}^n \pi_j \mathcal{N}(x \mid \mu_j, \sigma_j)$$

- $\sum_{j=1}^n \pi_j = 1$

- Nun sollen die Parameter  $\mu_j, \sigma_j, \pi_j, (1 \leq j \leq n)$  aus den Daten geschätzt werden.
- Mit der log-Likelihood Methode ergibt sich:

$$\sum_{i=1}^N \log \sum_{j=1}^n \pi_j \mathcal{N}(x_i | \mu_j, \sigma_j) \rightarrow \max$$

### Wenn die Daten nicht einer bekannten Verteilung folgen?

- Dichte basierend auf Histogramm
- Die Punkte stammen aus einer unbekannten Verteilung mit Dichte  $p(\cdot)$
- Die Punkte sind im 1-dimensionalen Euklidischen Raum eingebettet
- $p(x)$  soll geschätzt werden. Dazu wird ein kleines Intervall  $R$  um  $x$  betrachtet.
- Bilde  $P = \int_R p(x) dx$  und ziehe Stichprobe der Größe  $N$  aus  $p(x)$
- Jeder Punkt hat eine Wahrscheinlichkeit von  $P$  in  $R$  zu fallen.
- Die Zahl  $K$  der Punkte, die in  $R$  fallen, ist binomialverteilt

$$\text{Bin}(K | N, P) = \binom{N}{K} p^K (1 - P)^{N-K}$$

- Die Varianz und der Erwartungswert der Binomialverteilung sind bekannt:
  - Der erwartete Anteil der Punkte, die in  $R$  fallen, ist  $P$ .
  - Für die Varianz um diesen Mittelwert  $\frac{P(1-P)}{N}$
- Dadurch kann festgestellt werden, dass:
  - Für große  $N$  hat die Verteilung einen scharfen Peak um das erwartete  $P$  und damit  $K \approx NP$ .
  - Wird das Intervall ausreichend klein gemacht, so dass  $p(x)$  annähernd konstant in  $R$  ist, so erhält man  $P \approx p(x)V$  mit  $V$  ist "Volumen" von  $R$ .
  - Kombinieren beider Beobachtungen führt zu:

$$p(x) = \frac{K}{NV}$$

- $K$  kann nun fixiert und  $V$  aus den Daten ermittelt werden ( $k$ -NN)
- Alternativ kann  $V$  fixiert und  $K$  aus den Daten ermittelt werden (Kernel-Technik).
- Betrachte zunächst die  $k$ -NN Methode:
  - $p(x)$  soll an  $x$  geschätzt werden. Wähle  $K$  fest, aber beliebig.

- Lege Intervall um  $x$  und lasse es wachsen, bis es genau  $K$  Punkte umfasst.
- $p(x) = \frac{K}{NV}$  mit  $N$  der Stichproben
- Das Kernelverfahren verwendet eine Kernel-Funktion:

$$\kappa(u) = \begin{cases} 1 & |u| \leq \frac{1}{2} \\ 0 & \text{sonst} \end{cases}$$

- $\kappa$  repräsentiert ein Einheitsintervall (Parzen-Fenster).
- $\kappa(\frac{x-x_i}{h})$  ist 1, falls der Punkt  $x_i$  in einem Intervall der Länge  $h$  und Mitte  $x$  liegt.
- Die Gesamtanzahl ist somit

$$K_x = \sum_{i=1}^N \kappa\left(\frac{x-x_i}{h}\right)$$

- Es ist bereits bekannt, dass  $p(x) = \frac{K}{NV}$  und es ergibt sich damit

$$p(x) = \frac{1}{Nh} \sum_{t=1}^N \kappa\left(\frac{x-x_i}{h}\right)$$

mit  $V = h$ .

- Es kann auch ein Gauß-Kern verwendet werden, um zu einer stetigen Dichte zu gelangen:

$$p(x) = \frac{1}{N} \sum_{t=1}^N \frac{1}{\sqrt{w\pi h^2}} \cdot e^{-\frac{\|x-x_i\|^2}{2h^2}}$$

- Dabei repräsentiert  $h$  die Standardabweichung der Gauß-Komponente.

### 5.1.7 Satz von Bayes – im multivariaten Fall

- Es bleibt vieles beim Alten:

$$P(C | x) = \frac{P(C)p(x | C)}{p(x)}$$

- Die Dichten  $p(x | C)$  und  $p(x)$  sind nun aber multivariat.
  - Der ML Schätzer funktioniert wie gewohnt; es gibt jedoch mehr Parameter zu schätzen.
  - Nach wie vor können Mixtures, z.B. Mixture of Gaussians verwendet werden.

## Multivariate Normalverteilung

- $p(x \mid \Sigma, \mu) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}}$
- Mit den ML Schätzern:

$$m_j = \frac{\sum_{i=1}^N x_{ij}}{N}$$
$$s_{ij} = \frac{\sum_{k=1}^N (x_{ki} - m_i)(x_{kj} - m_j)}{N}$$

- Die Kernel- und  $k$ -NN Schätzer müssen nur geringfügig angepasst werden:
  - Anstatt eines Intervalls, wird eine Hyperkugel um  $x$  gelegt; die Kugel wächst, bis sie genau  $K$  Daten enthält.
  - Im Fall des Kernels repräsentiert das Parzen Fenster nun kein Intervall mehr, sondern eine Hyperkugel.
- Es können jedoch auch vereinfachende Annahmen gemacht werden (Naive Bayes).
- Eine (naive) Möglichkeit zur Vereinfachung ist anzunehmen, dass für jede Klasse  $C_k$  die Verteilungen der individuellen Attribute bedingt unabhängig sind; dann gilt:

$$p(x_i \mid C_k) = \prod_{1 \leq l \leq d} p(x_{il} \mid C_k)$$

- Somit ergibt sich final:

$$P(C_k \mid x_i) = \frac{P(C_k) \prod_{l=1}^d p(x_{il} \mid C_k)}{\sum_j P(C_j) \prod_{l=1}^d p(x_{il} \mid C_j)}$$

## 5.2 Gradientenabstieg

- Das Mixture of Gaussians Modell konnte nicht analytisch gelöst werden.
- Es kann jedoch der Gradient berechnet werden.
- Annahme:  $f(x)$  ist zu minimieren und der Gradient  $\nabla f(f)$  ist bekannt.

## 5.3 Exponentielle Verteilungen

- Gegeben durch

$$p(x \mid \eta) = h(x)g(\eta)e^{\eta^T u(x)}$$

mit

$$g(\eta) \int h(x)e^{\eta^T u(x)} dx = 1$$

- $x$  kann Skalar oder Vektor sein (sowohl diskret oder kontinuierlich).
- $\eta$  sind sogenannte *natürliche* Parameter der Verteilung.
- $u(x)$  ist eine Funktion von  $x$ .

### 5.3.1 Bernoulli Verteilung

- Ist bekanntlich gegeben durch:

$$p(x | \mu) = \mu^x (1 - \mu)^{1-x}$$

- Rechts wird eine Potenz des Logarithmus verwendet

$$\begin{aligned} p(x | \mu) &= e^{x \ln \mu + (1-x) \ln(1-\mu)} \\ &= (1 - \mu) e^{\ln(\frac{\mu}{1-\mu})x} \end{aligned}$$

und erkennen, dass

$$\begin{aligned} \eta &= \ln\left(\frac{\mu}{1-\mu}\right) \\ u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= \sigma(-\eta) \end{aligned}$$

mit

$$\sigma(\eta) = (1 + e^{-\eta})^{-1}$$

- Damit gilt:

$$p(x | \mu) = \sigma(-\eta) e^{\eta x}$$

### 5.3.2 Normalverteilung

- Ist bekanntlich gegeben durch:

$$\begin{aligned} p(x | \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{2\sigma^2}(x-\mu)^2} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2)} \end{aligned}$$

- Betrachte sofort das Ergebnis mit  $\eta = (\eta_1, \eta_2)$  und

$$\begin{aligned} \eta &= \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \\ u(x) &= \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ h(x) &= (2\pi)^{-\frac{1}{2}} \\ g(\eta) &= (-2\eta_2)^{\frac{1}{2}} e^{\frac{\eta_1^2}{4\eta_2}} \end{aligned}$$

## 5.4 Ausreichende Statistik

- Es kann auch der Likelihood berechnet werden:

$$p(X | \eta) = \left( \prod_{i=1}^N h(x_i) \right) g(\eta)^N e^{\eta^T \sum_{i=1}^N u(x_i)}$$

- Setzen des Gradienten von  $\ln(p(X | \eta))$  bzgl.  $\eta$  gleich Null liefert:

$$-\nabla \ln g(\eta_{ML}) = \frac{1}{N} \sum_{i=1}^N u(x_i)$$

- Für die Bernoulli-Verteilung genügt es die Summe der  $x$  zu speichern.
- Für die Normalverteilung wird die Summe von  $x$  und  $x^2$  benötigt.
- Für  $N \rightarrow \infty$  ergibt sich  $E(u(x))$  und  $\eta_{ML}$  geht in das wahre  $\eta$  über.

## 6 Clustering

## 7 Warenkorbanalyse

## 8 Analyse von Graphdaten