

Maschinelles Lernen

Zusammenfassung

Thomas Mohr

Contents

1	Grundlagen	3
1.1	(Un)-überwachtes Lernen	3
1.2	Inkrementelles Lernen	3
1.3	Aktives Lernen	3
1.4	Data cleansing	3
1.5	Datensatz	3
2	Deskriptive Statistik	4
2.1	Beschreibung von Daten	4
2.2	Mittelwert	4
2.3	Median und Midrange	5
2.4	Modus	5
2.5	Varianz und Schiefe	5
2.5.1	Moment k -ter Ordnung	6
2.6	Quantil	6
2.6.1	Interquantile range (IQR)	7
2.7	Korrelation zwischen Attributen	7
2.7.1	Kovarianz für numerische Daten	7
2.7.2	Korrelationskoeffizienten für numerische Daten	8
2.7.3	Rangkorrelationskoeffizient	8
2.7.4	χ^2 -Test	8
2.8	Visualisierung	9
2.8.1	Boxplots	9
2.8.2	Histogramme	9
2.8.3	Quantil-Plots	9
2.9	Distanzen	10
2.9.1	Distanz auf numerischen Attributen	10
2.9.2	Distanz auf ordinalen Attributen	10
2.9.3	Distanz auf nominalen Attributen	11
2.9.4	Distanz auf binären Attributen	11
2.9.5	Distanz auf gemischten Typen	11
3	Regression	12
4	Klassifikation	12
5	Clustering	12
6	Warenkorbanalyse	12
7	Analyse von Graphdaten	12

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})
 - **Klassifikation** im Fall von diskreten Labeln (z.B. *TRUE*, *FALSE*; ausgezeichnet, durchschnittlich, schlecht)
- 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 μ_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 σ symbolisiert

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

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

- $v(X_j) < 0$: 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

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

- $w(X) < 0$: 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
 - χ^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 $[-1, 1]$

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

- 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

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

mit $\mu\text{Rank}(X_i)$ 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 χ^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 χ^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 χ^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

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 $\delta_{if}^{(f)} = 1$.

- 3 Regression**
- 4 Klassifikation**
- 5 Clustering**
- 6 Warenkorbanalyse**
- 7 Analyse von Graphdaten**