4. Induktive Inferenz

Literatur:

Jorma Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific 1989.

Ming Li, Paul Vitányi, An Introduction to Kolmogorov Complexity and Its Applications, 2nd edn., Springer 2002.

S. Jain, D. Osherson, J.S. Royer, A. Sharma, Systems That Learn: An Introduction to Learning Theory, 2nd edn., MIT Press 1999.

C.S. Wallace, Statistical and Inductive Inference by Minimum Message Length, Springer 2005.

P.D. Grünwald, I.J. Myung, M.A. Pitt, eds. Advances in Minimum Description Length: Theory and Applications, MIT Press 2005.

K.P. Burnham, D.R. Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. Springer 2002.

experimentelle Praxis

⇝

Problem:

Ableitung von allgemeinen Gesetzmäßigkeiten aus beobachteten Daten. Diese sollen zum einen die beobachteten Daten erklären und zum anderen Vorhersagen über die Daten, die man bei der Durchführung von weiteren Experimenten erhalten würde, ermöglichen.

Häufig sind mehrere Erklärungen der beobachteten Daten möglich.

⇝

Frage:

Welche der möglichen Erklärungen soll ausgewählt werden?

- Prinzip der mehrfachen Erklärung lässt alle Theorien, die mit den beobachteten Daten konsistent sind, als Erklärung zu.

  Vorteil:

  Auch nach Erhalt von weiteren Daten befinden sich alle mit den beobachteten Daten konsistenten Theorien unter den zugelassenen Theorien.

  Nachteil: äußerst aufwendig

- Ockham's Rasiermesser-Prinzip wählt unter denjenigen Theorien, die mit den beobachteten Daten konsistent sind, die einfachste Theorie aus.

  ⇝

  Frage: Welche ist die einfachste Theorie?

  Nachteil:
  Nach Erhalt von weiteren Daten könnte die gewählte Theorie nicht mehr konsistent mit den beobachteten Daten sein.

Folgende Regel hat bei der Beantwortung obiger Frage eine probabilistische Sicht:

Seien D die beobachteten Daten, H eine Hypothese zur Erklärung der Daten D und $P(H)$ eine anfängliche Schätzung der Wahrscheinlichkeit, dass die Hypothese H zutreffend ist.

- Bayes'sche Regel
  Die Wahrscheinlichkeit, dass die Hypothese H zutrifft, ist proportional zum Produkt von $P(H)$ mit der durch die Hypothese H bedingten Wahrscheinlichkeit der beobachteten Daten D.

  ⇝

  Frage: Wie erhält man die anfängliche Schätzung der Wahrscheinlichkeiten?

Nach Beobachtung von neuen Daten werden die Wahrscheinlichkeiten der einzelnen Hypothesen aktualisiert. Ist die Wahrscheinlichkeit der neuen Daten für eine gegebene Hypothese klein, dann verringert die Multiplikation der bedingten Wahrscheinlichkeit mit der bisherigen Schätzung für P(H) den Wert von P(H) entsprechend.

Ziel:
Konstruktion eines universellen induktiven Inferenzsystems.

Dabei soll keine Theorie, die zu den beobachteten Daten konsistent ist, ausgeschlossen und unter den betrachteten Theorien die einfachen Theorien bevorzugt werden. Nach Erhalt von neuen Daten soll der Wissensstand mit Hilfe der Bayes'schen Regel aktualisiert werden.

4.1 Ein universelles induktives Inferenzsystem

gewünschte Eigenschaften:

i) Das universelle induktive Inferenzsystem sollte in der Lage sein, jede berechenbare deterministische oder stochastische Hypothese zu behandeln.

ii) Es sollte jede Art von Vorwissen verarbeiten

können. Dieses kann sich erstrecken von überhaupt kein Vorwissen bis zum Kennen der richtigen Hypothese.

iii) Zu jedem Zeitpunkt besitzt das induktive Inferenzsystem ein aktuelles Wissen bezüglich der Gültigkeit der einzelnen Hypothesen. In Abhängigkeit von neuen beobachteten Daten aktualisiert das induktive Inferenzsystem dieses Wissen.

↪

Fragen:

1) Wie spezifiziert das universelle induktive Inferenzsystem beobachtete Daten und Hypothesen?

2) Wie spezifiziert das induktive Inferenzsystem das aktuelle Wissen bezüglich der Gültigkeit der einzelnen Hypothesen und wie wird dieses Wissen aufgrund neuer beobachteter Daten aktualisiert?

Beobachtung:

1) Beobachtete Daten können unabhängig davon, wie das Experiment aussieht, binär kodiert werden. Im Prinzip könnten unendlich viele Experimente erfolgen, was dann in einem

unendlichen Binärstring führen würde.

$\Rightarrow$

Die bisher beobachteten Daten sind stets Präfix von möglichen unendlichen Binärstrings

Um die in dem unendlichen Binärstring enthaltenen Regelmäßigkeiten zu beschreiben, versucht das induktive Inferenzsystem auf der Basis des gegenwärtigen Wissens und der zuletzt beobachteten Daten eine Theorie zu entwickeln.

$\rightsquigarrow$

Hypothesen von möglichen Theorien werden identifiziert mit Algorithmen, die Binärstrings aus Ω, die den aktuellen Datenstring als Präfix enthalten, berechnen.

Es gibt unendlich viele Möglichkeiten, einen endlichen Präfix aus $\{0,1\}^*$ unendlich fortzusetzen. Die Strategie, gemäß der der aktuelle Präfix fortgesetzt wird, kann deterministisch sein, muss aber nicht. Diese kann selbst stochastisch sein. Dies bedeutet insbesondere, dass die Strategie den aktuellen Präfix auf unterschiedliche Art und Weise fortsetzen kann, wobei manche Fortsetzungen wahrscheinlicher sein können als andere.

$\Longrightarrow$

Es ist sinnvoll, jede Hypothese mit einer probabilistischen Maschine ohne Eingabe zu identifizieren

Satz 3.11 impliziert, dass äquivalent hierzu jede Hypothese mit einem von unten aufzählbaren Semimaß auf $\mathbb{N}$ identifiziert werden kann.

2) Das aktuelle Wissen bezüglich der Gültigkeit der einzelnen Hypothesen kann als Wahrscheinlichkeit, dass betreffende Hypothese zutrifft, interpretiert werden.

Struktur des universellen induktive Inferenzsystems: (UIIS)

Das UIIS enthält die im Beweis von Satz 3.12 konstruierte Maschine M.

$\Rightarrow$

UIIS kann jede probabilistische Maschine ohne Eingabe simulieren. Das aktuelle Wissen wird durch die Wahrscheinlichkeiten p_i, $i \in \mathbb{N}_0$, mit denen M die Maschine M_i auswählt, beschrieben.

Frage:
Was tun, wenn über die Gültigkeit der einzelnen Hypothesen kein Vorwissen vorliegt?

Antwort:

Das UIIS wird mit geeigneten a priori Wahrscheinlichkeiten gestartet. Diese geben uns für jede Hypothese eine anfängliche Schätzung der Wahrscheinlichkeit, dass diese Hypothese zutrifft.

Nach Erhalt neuer Daten werden die Wahrscheinlichkeiten p_i, $i \in \mathbb{N}_0$ unter Anwendung der Bayes'schen Regel aktualisiert. Die aktualisierten Wahrscheinlichkeiten heißen a posteriori Wahrscheinlichkeiten.

Fragen:

1) Was ist eine geeignete Wahl für die a priori Wahrscheinlichkeiten?

2) Wie sieht nach Erhalt von neuen Daten konkret die Anwendung der Bayes'schen Regel aus?

3) Wie schätzt man bei gegebenen bisher beobachteten Datenstring x das nächste Datensymbol? (Inferenzproblem)

Bevor wir die obigen Fragen beantworten, betrachten wir die Anwendung der Bayes'schen Regel genauer.

Sei S ein Ereignisraum von höchstens abzählbar vielen paarweise disjunkten Ereignissen.

Sei $H_0, H_1, H_2, \ldots$ eine Aufzählung von abzähl-

bar vielen Hypothesen betreffend eines betrachteten Phänomens. D.h., jedes H_i ist eine Wahrscheinlichkeitsverteilung über S.

$\mathcal{H} := \{H_0, H_1, H_2, \ldots\}$ heißt Hypothesenraum

Seien

- $\mathcal{H}$ erschöpfend, d.h., mindestens eine Hypothese trifft zu und gegenseitig ausschließend, d.h., maximal eine der Hypothesen trifft zu.
- D die beobachteten Daten eines Experimentes bezüglich des betrachteten Phänomens
- P eine gegebene a priori Wahrscheinlichkeitsverteilung auf $\mathcal{H}$.

Dann gilt: $\sum_i P(H_i) = 1$.

Annahme:
$\forall H_i \in \mathcal{H}$ können wir die bedingte Wahrscheinlichkeit $P_r(D/H_i)$ der beobachteten Daten D unter der Annahme, dass H_i zutrifft, berechnen.

$\Rightarrow$

Wir können auch

$$P_r(D) := \sum_j P_r(D/H_j) \cdot P(H_j)$$

berechnen bzw., falls die Anzahl der Hypothesen

mit Wahrscheinlichkeit > 0 unendlich ist, ε-approximieren.

Aus der Definition der bedingten Wahrscheinlichkeit kann leicht folgende Bayes'sche Regel hergeleitet werden.

$$Pr(H_i|D) = \frac{Pr(D|H_i)\,P(H_i)}{Pr(D)}$$

$$= \frac{Pr(D|H_i)\,P(H_i)}{\sum_j Pr(D|H_j)\,P(H_j)}.$$

- $Pr(H_i|D)$ ist die aufgrund der Daten D aktualisierte Wahrscheinlichkeit der Hypothese H_i und demzufolge die a priori Wahrscheinlichkeit der Hypothese H_i für die nachfolgenden Daten.
- Die Wahrscheinlichkeit $Pr(D)$ der Daten D dient in obiger Gleichung als Normalisator, so dass die a posteriori Wahrscheinlichkeiten wieder eine Wahrscheinlichkeitsverteilung bilden, d.h., $\sum_i Pr(H_i|D) = 1$.

Ziel:
Beantwortung der obigen Fragen.

1) Wahl der a priori Wahrscheinlichkeiten

- Falls Vorwissen bezüglich der Wahrscheinlichkeiten der einzelnen Hypothesen vorhanden ist, dann kann die daraus resultierende Wahrs-

scheinlichkeitsverteilung als a priori Wahr-
scheinlichkeitsverteilung genommen werden.

Annahme: Es liegt kein Vorwissen vor.

Im Beweis von Satz 3.12 haben wir eine probabilistische Maschine M ohne Eingabe konstruiert, die

- eine Aufzählung $M_0, M_1, M_2, \ldots$ aller probabilistischen Maschinen ohne Eingabe vorgenommen und zufällig $i \in \mathbb{N}_0$ gewählt hat. Dabei ist p_i, $i \in \mathbb{N}_0$ die Wahrscheinlichkeit, dass i gewählt wird.
- Danach simuliert M die Maschine M_i.

Für alle $i \in \mathbb{N}_0$ ist dann das universelle von unten aufzählbare Semimaß $m(i)$ definiert durch

$$m(i) := p_i \mu(M_i),$$

wobei $\mu(M_i)$ das zur Maschine M_i korrespondierende von unten aufzählbare Semimaß bezeichnet.
Die einzige Voraussetzung bezüglich der Wahrscheinlichkeiten p_i ist $p_i > 0 \ \forall i \in \mathbb{N}_0$.

Idee:
Wähle die Wahrscheinlichkeiten p_i, $i \in \mathbb{N}_0$ geeignet und definiere mit Hilfe dieser die a priori Wahrscheinlichkeit für jedes von unten aufzählbare

Semimaß m'.

Bemerkung:
Es können $i, j \in \mathbb{N}_0$ mit $i \neq j$ und $\mu(M_i) = \mu(M_j)$ existieren. Daher gilt für die Wahrscheinlichkeit $P_M(m')$, dass die Maschine M das von unten aufzählbare Semimaß m' wählt

$$P_M(m') = \sum_{i : \mu(M_i) = m'} p_i .$$

Für $x \in \{0,1\}^*$ definieren wir

$$m'(x) = m'(\Omega_x).$$

Da m' ein Semimaß ist, gilt

$$m'(x) \geq m'(x0) + m'(x1).$$

Für $y \in \{0,1\}^*$ definieren wir das bedingte Semimaß $m'(y|x)$ durch

$$m'(y|x) := \frac{m'(xy)}{m'(x)} .$$

Übung:
Sei m' ein Semimaß auf Ω. Zeigen Sie, dass dann auch $m'(\cdot|x)$ für alle $x \in \{0,1\}^*$ ein Semimaß auf Ω_x ist.

Ziel:
Definition einer geeigneten a priori Wahrscheinlichkeit $P_M(m')$ für jedes von unten aufzählbare

Seminaß m'.

gewünschte Eigenschaft:
"Einfachere" Seminaße sollen bevorzugt werden.
D.h., einer probabilistischen Maschine ohne Eingabe, deren Programm eine kleinere Größe hat, soll eine größere Wahrscheinlichkeit zugeordnet werden als einer probabilistischen Maschine ohne Eingabe mit größerer Programmgröße.

Für $i \in \mathbb{N}_0$ sei k_i diejenige Kodierung der probabilistischen Maschine M_i, die die Maschine M aufzählt. Wir definieren dann die Wahrscheinlichkeit p_i durch

$$p_i := 2^{-|k_i|}.$$

Bemerkung:
Damit die Summe $\sum_i 2^{-|k_i|}$ nicht größer als eins wird, benötigen wir folgende Eigenschaft:

- Für $i \neq j$ ist weder k_i ein Präfix von k_j noch k_j ein Präfix von k_i.

Dann können wir genauso wie im Beweis von Satz 3.7 $\sum_i 2^{-|k_i|} \leq 1$ beweisen.

Hierzu muss M noch geeignet modifiziert werden.

Übung
Modifizieren Sie die Maschine M, so dass diese obige Eigenschaft besitzt.

$\Rightarrow$

$$P_M(m') := \sum_{i:\, M(M_i)=m'} p_i = \sum_{i:\, M(M_i)=m'} 2^{-|\xi_i|}.$$

Beobachtung:
Für jedes von unten aufzählbare Semimaß m' existiert in der Aufzählung $M_0, M_1, M_2, \ldots$ der probabilistischen Maschinen ohne Eingabe mindestens eine Maschine M_j mit $\mu(M_j) = m'$.

$\Rightarrow$

$P_M(m') > 0$ $\forall$ von unten aufzählbare Semimaße m'.

Ziel:
Lösung des Inferenzproblems.

Sei x der bisher beobachtete Datenstring.
Für $i \in \mathbb{N}_0$ bezeichne $\bar{p}_i$ die aktuelle Wahrscheinlichkeit, dass M die Maschine M_i für die Simulation wählt.

Zur Lösung des Inferenzproblems benötigen wir die bedingten Wahrscheinlichkeiten $P_T(0|x)$ und $P_T(1|x)$. Für diese gilt:

$$Pr(0|x) = \sum_i \bar{p}_i \cdot \mu(M_i)(0|x)$$

$$= \sum_i \bar{p}_i \cdot \frac{\mu(M_i)(x0)}{\mu(M_i)(x)}$$

und analog

$$Pr(1|x) = \sum_i \bar{p}_i \frac{\mu(M_i)(x1)}{\mu(M_i)(x)}.$$

Als nächstes Symbol wird dasjenige $y \in \{0,1\}$ geschätzt, das die maximale bedingte Wahrscheinlichkeit hat.

Übung:
Seien x die bisher verarbeiteten Daten, y die neuen beobachteten Daten und $\bar{p}_i$, $i \in \mathbb{N}_0$ die aktuellen Wahrscheinlichkeiten, dass M die Maschine M_i für die Simulation auswählt. Führen Sie die durch y bedingte Aktualisierung der Wahrscheinlichkeiten durch.

Bemerkung:
Die oben definierten a priori Wahrscheinlichkeiten des universellen induktiven Inferenzsystems bevorzugen probabilistische Maschinen ohne Eingabe mit kurzer Kodierung. Aufgrund der Struktur der Bayes'schen Regel kann dies nur durch die bedingten Wahrscheinlichkeiten der beobachteten Daten geändert werden. Bei gleichen bedingten Wahrscheinlichkeiten behalten Maschinen mit

keiner Kodierung ihre Bevorzugung.

Das universelle induktive Inferenzsystem kann nicht direkt in ein praktisches Inferenzsystem umgesetzt werden. Bei der Entwicklung eines praktischen Inferenzsystem haben wir dieselben Fragen wie bei der Entwicklung des universellen Inferenzsystems zu beantworten.

1) Wie spezifiziert das induktive Inferenzsystem beobachtete Daten und Hypothesen?

2) Wie spezifiziert das induktive Inferenzsystem das aktuelle Wissen bezüglich der Gültigkeit der einzelnen Hypothesen und wie wird dieses Wissen aufgrund neuer beobachteter Daten aktualisiert?

Während wir bei der Entwicklung des universellen induktiven Inferenzsystems die Abstraktion so weit wie möglich betrieben haben, ist bei der Entwicklung von praktischen induktiven Inferenzsystemen die konkrete Situation zu berücksichtigen. Um ein Gefühl dafür zu bekommen, wie dies bewerkstelligt werden kann, werden wir zunächst konkretere theoretische Systeme skizzieren.

4.2 Die Identifizierung von Sprachen

Literatur:

D. Angluin, C.H. Smith, Inductive Inference: Theory and Methods, Computing Surveys 15 (1983), 237 - 269.

S. Jain, D. Osherson, J.S. Royer, A. Sharma, Systems That Learn: An Introduction to Learning Theory, 2nd edn., MIT Press 1999.

Ziel:

Konkretisierung eines präzisen Modells für empirische Untersuchungen.

Um ein induktives Inferenzproblem zu definieren sind folgende fünf Konzepte zu spezifizieren:

1) Eine theoretisch mögliche Realität,
2) verständliche Hypothesen,
3) verfügbare Daten bezüglich jeder gegebenen Realität,
4) ein Inferenzsystem und
5) Kriterien für eine erfolgreiche Inferenz.

In unserem universellen Inferenzsystem waren mögliche Realitäten Binärstrings aus Σ. Hypothesen entsprachen Algorithmen, die Binärstrings aus Σ,

die den aktuellen Datenstring als Präfix enthalten, berechnen. Verfügbare Daten waren Elemente aus $\{0,1\}^*$.

Folgende wichtige Frage wurde in der Lerntheorie intensiv untersucht:

Für welche Klassen von möglichen Realitäten existieren Inferenzsysteme, die bezüglich einer beliebigen Realität aus dieser Klasse garantiert Erfolg haben?

Viele Realitäten lassen sich als Sprachen modellieren

$\Rightarrow$

Die Untersuchung der Identifikation von Sprachen ist interessant.

Sei Σ ein endliches Alphabet. Eine <u>Sprache über Σ</u> ist eine Teilmenge von Σ^*. Eine <u>Sprachklasse</u> ist eine Kollektion von Sprachen über Σ.

Die Identifikation einer Sprache als induktives Inferenzproblem könnte wie folgt spezifiziert werden:

1) Als theoretisch mögliche Realität könnte eine Sprachklasse über ein Alphabet Σ dienen.

   <u>Bsp.:</u>
   a) Klasse der regulären Sprachen über $\{0,1\}$.
   b) Alle Sprachen über $\{0,1\}$.

2) Generatoren oder Akzeptoren von Sprachen könnten als Hypothesen dienen

Bsp.:

a) reguläre Ausdrücke, reguläre Grammatiken oder endliche Automaten

b) allgemeine Grammatiken oder Turing-maschinen.

3) Annahme:

Eine unbekannte Sprache L soll identifiziert werden.

Verfügbare Daten können dergestalt sein, dass

- nur positive Information, d.h., Elemente von L gegeben werden

oder dass

- sowohl positive als auch negative Information gegeben wird.

Eine positive Präsentation von L ist eine Aufzählung der Elemente von L. Eine vollständige Präsentation ist eine Aufzählung von Paaren $\langle s, d \rangle \in \Sigma^* \times \{0,1\}$, wobei

i) $$d = \begin{cases} 1 & \text{falls } s \in L \\ 0 & \text{sonst} \end{cases}$$

und

ii) jedes $s \in \Sigma^*$ als erste Komponente eines Paares in der Aufzählung vorkommt.

Einen anderen Zugang erhält man, wenn der Inferenzsystem Elemente $s \in \Sigma^*$ generieren könnte und ein Orakel dem Inferenzsystem mitteilt, ob $s \in L$ oder nicht. Eine derartige Präsentation von L heißt <u>Präsentation durch einen Informanten</u>.

4) Ein Inferenzsystem ist dann ein Algorithmus, der aufgrund der Daten Hypothesen aufstellt. Wie dieser Algorithmus konkret arbeitet, hängt von der Präsentation der zu identifizierenden Sprache L ab. Nach Erhalt von neuen Daten kann der Algorithmus von der bisherigen aktuellen Hypothese zu einer neuen Hypothese wechseln.

5) Zwei klassische Konzepte für erfolgreiche Inferenz sind

i) Identifikation im Grenzwert und
ii) Identifikation durch Aufzählung.

Wir werden nun beide Konzepte skizzieren.

i) <u>Identifikation im Grenzwert:</u>

Die Identifikation im Grenzwert betrachtet induktive Inferenz als einen unendlichen Prozess.

Annahme:

M ist eine induktive Inferenzmethode, die versucht korrekt eine unbekannte Sprache L zu identifizieren.

- M erhält eine ständig größer werdende Kollektion von Daten bezüglich L und generiert eine unendliche Folge

$$g_1, g_2, g_3, \ldots$$

von Hypothesen.

Falls $m \in \mathbb{N}$ existiert, so dass

a) g_m ist eine korrekte Beschreibung von L und

b) $g_m = g_{m+1} = g_{m+2} = \ldots$,

dann sagen wir, dass M die Sprache L korrekt im Grenzwert identifiziert.

Bemerkung:

Da stets neue Daten mit der aktuellen Hypothese nicht konsistent sein könnten, kann M nicht entscheiden, wann es zu einer korrekten Hypothese konvergiert ist.

ii) Identifikation durch Aufzählung

Unter Identifikation durch Aufzählung versteht man

die systematische Durchmusterung des Hypothesen-raumes bis eine Hypothese gefunden ist, die mit allen bisherigen Daten konsistent ist.

Annahme:

1) Eine korrekte Hypothese ist stets zu allen verfügbaren Daten konsistent.

2) Eine falsche Hypothese ist zu einer genügend großen Kollektion von Daten inkonsistent.

Identifikation durch Aufzählung setzt voraus, dass der Hypothesenraum aufgezählt werden kann. Ist dies der Fall und sind obige Annahmen erfüllt, dann garantiert Identifikation durch Aufzählung die Identifikation im Grenzwert.

Ein induktives Inferenzsystem, das Identifikation im Grenzwert kann nicht entscheiden, wann die korrekte Hypothese erreicht ist. In der Praxis benötigt man Kriterien, die etwas über die Qualität der aktuellen Hypothese aussagt.

Ziel:
Diskussion einiger solcher Kriterien.

Zwei Kriterien bezüglich der Bewertung von Hypothesen liegen auf der Hand:

i) Die "Einfachheit" der Hypothese.
ii) Die Relation der Hypothese zu den verfügbaren Daten.

Bezüglich derartiger Kriterien können wir eine Ordnung der Hypothesen bezüglich der betrachteten Datenmenge S definieren.

Seien g und h zwei Hypothesen. Wir schreiben

$$g \leq_S h,$$

falls die Hypothese g eine "bessere" Erklärung der Datenmenge S als die Hypothese h ergibt. Wenn eine derartige Aussage bezüglich zwei Hypothesen nicht möglich ist, dann sind diese nicht vergleichbar.

$\Rightarrow$

Die Relation $\leq_S$ definiert eine partielle Ordnung auf den Hypothesenraum. Die Menge der <u>besten Hypothesen bezüglich einer Datenmenge S</u> sind gerade die minimale Elemente in dieser partiellen Ordnung. Wir nennen eine solche partielle Ordnung <u>Güteordnung</u>. Die Menge

$$\{ \leq_S \mid S \text{ endliche Datenmenge} \}$$

definiert eine Familie von Güteordnungen.

Eine Familie von Güteordnungen heißt <u>berechenbar</u>,

falls es einen Algorithmus O_1 gibt, der eine beliebige endliche Datenmenge S als Eingabe erhält und eine bezüglich S beste Hypothese, also ein minimales Element der Ordnung $\leq_S$, ausgibt.

Eine Familie von Gütordnungen heißt datenunabhängig, falls $\forall$ Hypothesen g, h und alle endliche Datenmengen S

$$g \leq_S h \Rightarrow$$

$g \leq_{S'} h$ oder g und h unvergleichbar bzgl. S' für alle endliche Datenmengen S'.

Ordnungen, die nur auf Eigenschaften der Hypothesen wie z.B. ihre Größe abhängen, sind somit datenunabhängig.

Beispiel: (Einfachheit der Hypothesen)

Aufgabe:
Gegeben zwei disjunkte endliche Mengen S_0 und S_1 finde einen endlichen Automaten mit Minimalanzahl von Zuständen, der alle Strings in S_1 und keinen String in S_0 akzeptiert.

Dieses Problem ist zwar berechenbar, jedoch NP-hart. (Referenz siehe Übersichtsartikel von Angluin/Smith)

Forschungsprojekt:
"Praktische" Algorithmen zur Identifikation von Sprachen.

4.3. Induktive Inferenzsysteme unter Verwendung von MDL und MML

Ziel:
Entwicklung eines theoretischen induktiven Inferenzsystems, das etwas strukturierter als unser UIIS ist.

Idee:
Verwendung von zweigeteilten Kodes, wobei der erste Teil eine Theorie, die die Regelmäßigkeiten in den Daten erklärt, kodiert. Der zweite Teil ist dann eine Kodierung der Daten unter Verwendung des im ersten Teil beschriebenen Theorie.

Durchführung:

Gegeben seien die beobachteten Daten D und eine Menge H von Theorien zur Erklärung der Daten. Sowohl MDL (minimum description length) als auch MML (minimum message length) wählt diejenige Theorie aus, die die Summe

i) der Länge der Beschreibung der Theorie und

ii) der Länge der Beschreibung der Daten, wenn deren Kodierung mit Hilfe der kodierten Theorie erfolgt,

minimiert.

Bemerkung:

- Häufig sagt man anstatt Theorie auch Modell.
- Daten, die durch das Modell beschrieben werden und somit eine Kodierung mit Hilfe des Modells erlauben, heißen klassifiziert.

Frage:
Wie kann obiges informeller Ansatz formalisiert werden?

Idee:
Verwende analog zum UIIS die Bayes'sche Regel.

D.h., wir haben

- anfängliche Wahrscheinlichkeiten $P(H)$ für die Hypothesen $H \in \mathcal{H}$

und möchten in Abhängigkeit von den anfänglichen Wahrscheinlichkeiten und den beobachteten Daten

- die resultierenden Wahrscheinlichkeiten $Pr(H|D)$ für die Hypothesen $H \in \mathcal{H}$

berechnen.

$\Rightarrow$

Wir müssen uns überlegen, wie wir die anfängliche Wahrscheinlichkeiten erhalten.

Hierzu formulieren wir zunächst die Bayes'sche Regel äquivalent um. Es gilt

$$\Pr(H \mid D) = \frac{\Pr(D \mid H) \cdot P(H)}{\Pr(D)}$$

$$\Leftrightarrow -\log \Pr(H \mid D) = -\log \Pr(D \mid H) - \log P(H) + \log \Pr(D). \quad (*)$$

Gegeben die anfänglichen Wahrscheinlichkeiten P und die beobachteten Daten D besteht die Aufgabe darin, diejenige Hypothese $H_0 \in \mathcal{H}$ zu finden, die die bedingte Wahrscheinlichkeit $\Pr(H \mid D)$ maximiert. Wegen $(*)$ ist dies äquivalent dazu, diejenige Hypothese $H_0 \in \mathcal{H}$ zu finden, die $-\log \Pr(H \mid D)$ minimiert.

Da für gegebene Daten D und anfängliche Wahrscheinlichkeiten P die Wahrscheinlichkeit $\Pr(D)$ konstant ist, ist dies äquivalent zu folgender Vorgehensweise:

- Bestimme diejenige Hypothese $H_0 \in \mathcal{H}$, die

  $$-\log P(H) - \log \Pr(D \mid H)$$

  minimiert.

Bezeichne $Pr(H,D)$ die Wahrscheinlichkeit, dass H und D gemeinsam auftreten. Dann ist folgende Vorgehensweise zu obiger äquivalent:

- Bestimme diejenige Hypothese $H_0 \in \mathcal{H}$, die

$$Pr(H,D) = P(H) \cdot Pr(D|H)$$

maximiert.

Ab hier gehen MDL und MML unterschiedliche Wege. Wir werden zunächst MDL und dann MML beschreiben.

MDL:

Grundlage für den von MDL gegangenen Weg bildet die Präfixkomplexität. Satz 3.13 stellt einen Zusammenhang zwischen der Präfixkomplexität und einem universellen von unten aufzählbaren Semimaß her.

$\rightsquigarrow$

Idee:
Verwende ein universelles von unten aufzählbares Semimaß als apriori Wahrscheinlichkeit.

Zur Durchführung dieser Idee ist es notwendig, bedingte Wahrscheinlichkeit mit bedingter Präfixkomplexität in Beziehung zu setzen. Wir definieren die bedingte Präfixkomplexität $KP(x|y)$ von x, wenn y bekannt ist, analog zur bedingten

Kolmogorov-Komplexität:

$$KP(x|y) := K_U(x|y) = \min\{|p| \mid U(p,y) = x\},$$

wobei U ein asymptotisch optimaler Präfixdekompressionsalgorithmus ist.

Sei m ein universelles von unten aufzählbares Semimaß. Wenn wir nun als a priori Wahrscheinlichkeiten dieses Semimaß m verwenden, dann erhalten wir

$$\log P(H) := \log m(H) \quad \text{und}$$
$$\log P(D|H) := \log m(D|H).$$

Satz 3.13 $\Rightarrow$

$$-\log m(H) = KP(H) + d \quad \text{und}$$
$$-\log m(D|H) = KP(D|H) + d,$$

wobei d eine Konstante ist, die nicht von D und H abhängt.

Somit erhalten wir aus obiger Vorgehensweise folgendes Prinzip:

- Bestimme diejenige Hypothese $H_0 \in \mathcal{H}$, die

  $$KP(H) + KP(D|H)$$

  minimiert.

Die Präfixkomplexitätsfunktion KP ist nicht

berechenbar.

$\Rightarrow$

Das induktive Inferenzsystem muss für $H \in \mathcal{H}$ anstatt seine Präfixkomplexität eine einfache Approximation für $KP(H)$ verwenden.

Satz 3.8 $\Rightarrow$

$$KP(x) \leq 2|x| + c \quad \forall x \in \{0,1\}^*$$

Idee:
Approximiere $KP(H)$ und $KP(DIH)$ durch die daraus resultierenden oberen Schranken.

$\Rightarrow$

Wir benötigen Kodierungen von H und DIH als Strings über $\{0,1\}$.

Bezüglich der Kodierungen von H und DIH verwenden wir dann den Präfixkode, den wir durch Verdopplung der Bits mittels Anhängen von 01 erhalten.

Übung:
Wir haben die obere Schranke aus Satz 3.8 für $KP(x)$ in den Übungen verbessert. Nehmen wir an, dass wir anstatt obiger oberen Schranke für die Präfixkomplexität eines Strings x eine dieser besseren oberen Schranken verwenden. Inwiefern unterscheidet sich das resultierende Inferenzsystem von dem obigen?

Ziel: Konstruktion von Kodierungen für $\mathcal{H}$ und $D|H$.

Kodierung von $\mathcal{H}$:

- Falls wir einen Algorithmus $O_{\mathcal{H}}$ haben, der $\mathcal{H}$ aufzählt, dann können wir als Kode $\varphi(H)$ für H die Binärdarstellung desjenigen Index nehmen, den der Algorithmus $O_{\mathcal{H}}$ der Hypothese H zuordnet.
- Falls eine konkrete Kodierung aller Hypothesen $H \in \mathcal{H}$ vorliegt, dann kann der zur Kodierung von H korrespondierende Binärkode als $\varphi(H)$ genommen werden.

$\Rightarrow$

Wir erhalten eine Kodierung der Hypothesen in $\mathcal{H}$, indem wir einen Aufzählungsalgorithmus für $\mathcal{H}$ oder eine konkrete Kodierung aller Hypothesen in $\mathcal{H}$ konstruieren.

Kodierung von $D|H$:

Sei $D \in \{0,1\}^n$. Jede Hypothese in $\mathcal{H}$ definiert eine Wahrscheinlichkeitsverteilung auf $\{0,1\}^n$. Für $H \in \mathcal{H}$ bezeichne $P(\cdot|H)$ die korrespondierende Verteilungsfunktion. Bezüglich dieser Verteilungsfunktion können wir einen Huffman-Kode konstruieren. Für diesen gilt:

$$|\psi(D|H)| = -\log P_r(D|H).$$

Diesen Kode nehmen wir.

MML:

Während die Präfixkomplexität die Basis für die Überlegungen von MDL bildete ist die Shannon'sche Informationstheorie die Basis für die Überlegungen von MML.

Idee:
Interpretiere die Elemente des Hypothesenraumes $\mathcal{H}$ als Hypothesen über die Quelle der Daten. Konstruiere den Zusammenhang zwischen den Wahrscheinlichkeiten und der Kodierung der Hypothesen bzw. den bedingten Wahrscheinlichkeiten der Daten und der Kodierung der Daten unter der Annahme, dass die zugrunde liegende Hypothese H zutrifft direkt aus der Shannon'sche Informationstheorie.

$\rightsquigarrow$

Gegeben die a priori Wahrscheinlichkeiten $P(H)$, $H \in \mathcal{H}$ wird ein Kode φ gewählt, der die erwartete Länge der kodierten Hypothesen minimiert. Für die Kodierung der Daten unter der Annahme, dass die zugrunde gelegte Hypothese $H \in \mathcal{H}$ zutrifft, verwendet man einen optimalen Kode ψ.

$\Rightarrow$

Der Kode φ minimiert die erwartete Länge des kodierten Datenstrings unter der Annahme, dass die Hypothese H zutrifft. Gegeben die Wahrscheinlichkeitsverteilungen erhalten wir unter Verwendung des Huffmen-Kodes die Kodes φ und ψ.

Die Inferenzsysteme MDL und MML sind nahezu identisch. Im wesentlichen unterscheiden sie sich darin, wie sie die Kodierungen für die Hypothesen konstruieren. Während MML zunächst a priori Wahrscheinlichkeiten der Hypothesen und dann unter Verwendung der Huffmen-Kodierung seinen Kode für die Hypothesen konstruiert, erstellt MDL die Kodierung der Hypothesen direkt.

Sowohl bei der Verwendung von MDL als auch bei der Verwendung von MML hat man bei praktischen Anwendungen die Schwierigkeit, zunächst die Menge der möglichen Theorien oder Hypothesen zu modellieren und dann die a priori Wahrscheinlichkeitsverteilung direkt oder indirekt zu spezifizieren.

<u>Annahme</u>: Hypothesenraum $\mathcal{H}$ ist gegeben.

Falls kein Vorwissen vorliegt, dann ist es sinnvoll, jeder Hypothese $H \in \mathcal{H}$ dieselbe a priori Wahrscheinlichkeit $P(H)$ zu geben. D.h., falls

$|\mathcal{H}|$ endlich ist,

$$P(H) := \frac{1}{|\mathcal{H}|} \quad \forall H \in \mathcal{H}.$$

Dann kann bei der Berechnung einer Hypothese H mit maximaler a posteriori Wahrscheinlichkeit $Pr(H|D)$ der Term $P(H)$ herausgenommen werden, da dieser für alle Hypothesen gleich ist. Da $Pr(D)$ eine von H unabhängige Konstante ist, muss diese auch nicht berücksichtigt werden.

$\Rightarrow$

Eine Hypothese mit maximaler a posteriori Wahrscheinlichkeit $Pr(H|D)$ ist gemäß der Bayes'schen Regel eine, die

$$Pr(D|H)$$

maximiert.

$Pr(D|H)$ heißt <u>likelihood</u> der Daten, gegeben die Hypothese H. Eine Hypothese H, die $Pr(D|H)$ maximiert, heißt <u>maximum likelihood Hypothese</u>. D.h., H_{ml} ist eine maximum likelihood Hypothese, falls

$$Pr(D|H_{ml}) = \max_{H \in \mathcal{H}} Pr(D|H).$$

<u>Forschungsprojekt</u>:
"Praktische" Algorithmen zur Konstruktion von Hypothesenräume und a priori Wahrscheinlichkeits-

4.4 Eine Theorie der Ähnlichkeit

Sei S eine Menge von Objekten. Gegeben zwei beliebige Elemente x und y in S soll eine Aussage darüber gemacht werden, ob und inwieweit diese einander ähnlich sind.

Beispiel 4.1:

DNA-Sequenzen sind Strings über dem Alphabet $\Sigma := \{A, G, C, T\}$.

Sei S_1 eine Menge von DNA-Sequenzen von verschiedenen Spezies.

Evolutionäre Geschichte $\rightsquigarrow$ Sequenzen haben sich aufgrund von Mutationen geändert.

$\Rightarrow$

$x, y \in S_1$ haben in der evolutionären Geschichte gemeinsame Vorfahrsequenzen.

D.h., je ähnlicher x und y umso größer ist die Verwandtschaft der korrespondierenden Spezies

$\Rightarrow$

Es ist interessant, Aussagen über die Ähnlichkeit zweier DNA-Sequenzen zu machen.

□

Beispiel 4.2:

Sei S_2 eine Menge von Stühlen. S_2 enthält viele unterschiedliche Stuhltypen, z.B. vier-beinige Stühle, Freischwinger, Drehstühle u.s.w. Stühle können

- funktional oder auch
- geschmückt (z.B. kunstvoll gedrechselte Beine)

sein.

Gegeben zwei Stühle aus S_2 soll eine Aussage über ihre Ähnlichkeit gemacht werden.

Derartige Fragestellungen können auftreten, wenn wir

- die Menge der Stühle klassifizieren möchten oder
- ein Tool entwickeln wollen, das entscheidet, ob ein gegebenes Objekt ein Stuhl ist.

◊

↝

Frage:
Was ist ein sinnvolles Maß für die Ähnlichkeit zweier Objekte?

Idee:
Definiere eine Menge M von Operationen auf S, mittels derer Objekte in S verändert werden können und ordne diesen Operationen Kosten zu. Ein Maß für

die Ähnlichkeit zweier Objekte x und y könnten dann die Kosten einer kostengünstigsten Operationenfolge sein, die x nach y transformiert.

Beispiel 4.1 (Fortführung)

Operationen ≙

- Streichen einer Teilsequenz
- Einfügen einer Teilsequenz
- Ersetzung einer Teilsequenz durch eine andere

⋮

◊

Beispiel 4.2 (Fortführung)

Operationen ≙

- Glätten von Stuhlbeinen
- Anbringen von Drechseleien bei gegebenen Parametern
- Transformation einer Fläche in eine andere
- Vergrößern der Lehne unter vorgegebene Parametern

⋮

◊

Beispiel 4.2 zeigt, dass eine Operation r und die zugehörige inverse Operation r^{-1}, die die Operation r wieder rückgängig macht, nicht gleich schwer sein müssen.

⇒

Obige informelle Definition für das Maß der Ähnlich-

keit zweier Objekte ist nur sinnvoll, wenn die inverse Operation einer Operation stets genauso schwer ist wie die Operation selbst und daher dieselben Kosten hat.

Ziel:
Definition eines Maßes der Ähnlichkeit zweier Objekte bei gegebener Operationenmenge und gegebener Kostenfunktion für diese Operationenmenge.

Seien S eine Menge von Objekten und $S \subseteq S$ diejenige Objekte in S, für die wir uns konkret interessieren. Sei M eine Menge von Operationen, die Objekte in S in Objekte in S überführen. Sei

$$c: M \to \mathbb{Q}$$

eine Kostenfunktion, die jeder Operation in M Kosten zuordnet.

Für $r \in M$ schreiben wir genau dann $z_2 = r(z_1)$, $z_1, z_2 \in S$, wenn r angewandt auf z_1 das Objekt z_2 ergibt.

Seien x und y zwei beliebige verschiedene Objekte in S. Eine Operationenfolge $R := r_1, r_2, \ldots, r_t$ transformiert genau dann x nach y, wenn $z_1, z_2, \ldots, z_{t-1} \in S$ existieren, so dass

1) $z_1 = r_1(x)$,
2) $z_i = r_i(z_{i-1})$ für $2 \leq i \leq t-1$ und
3) $y = r_t(z_{t-1})$.

Wir schreiben dann auch $y = R(x)$.

Die Kosten $c(R)$ einer Folge $R := r_1, r_2, \ldots, r_t$ sind definiert durch

$$c(R) := \sum_{i=1}^{t} c(r_i).$$

Die Distanz $d_c(x,y)$ zweier Objekte x und y ergibt sich aus der hälftigen Summe der Kosten einer kostengünstigsten Operationenfolge R_1, die x nach y transformiert und einer kostengünstigsten Operationenfolge R_2, die y nach x transformiert. D.h.,

$$d_c(x,y) := \frac{1}{2}\Big(\min\{c(R_1) \mid y = R_1(x)\} + \min\{c(R_2) \mid x = R_2(y)\}\Big).$$

Falls keine Operationenfolge R_1 mit $y = R_1(x)$ oder keine Operationenfolge R_2 mit $x = R_2(y)$ existiert, dann vereinbaren wir $d_c(x,y) := \infty$.

Je kleiner die Distanz zweier Objekte, desto größer ist ihre Ähnlichkeit. Daher definieren wir die Ähnlichkeit $s_c(x,y)$ zweier Objekte x und y durch

$$s_c(x,y) := \frac{1}{d_c(x,y)}.$$

Dabei vereinbaren wir $s_c(x,y) = 0$, falls $d_c(x,y) = \infty$.

Frage:
Wie erhalten wir für eine gegebene Menge S von

Objekten eine geeignete Menge von Operationen?

In der Regel werden eine Vielzahl von Operationenmengen in Frage kommen. Welche von diesen wählen wir dann aus?

Idee:
Wende das allgemeine Prinzip, auf das MDL und MML aufbauen, hier an.

Durchführung:
Die Menge S von Objekten entspricht den beobachteten Daten und die Familie von möglichen Operationenmengen der Menge $\mathcal{H}$ von Theorien zur Erklärung der Daten

Anpassung des allgemeinen Prinzips:

Gegeben seien die Menge S von Objekten und eine Familie $\mathcal{M}$ von Operationenmengen. Gesucht ist eine Operationenmenge $M \in \mathcal{M}$, die die Summe

i) der Länge der Beschreibung der Operationenmenge und
ii) der Länge der Beschreibung der Objektmenge S, wenn deren Kodierung mit Hilfe der gewählten Operationenmenge M erfolgt,

minimiert.

Damit aus obigen Prinzip ein praktisches System entstehen kann, verlangen wir zusätzlich die Einhaltung der folgenden Regeln:

1. Jede Menge $M \in \mathcal{M}$ hat eine endliche Beschreibung.
2. Die Beschreibung der Operationenmenge M muss vollständig sein. D.h., sie beinhaltet auch für jede Operation den für die Durchführung der Operation benötigten Algorithmus.

Regel 1 $\Rightarrow$
Nur endlich viele verschiedene Typen von Operationen darf in einer Operationenmenge $M \in \mathcal{M}$ existieren. (Folgt eigentlich bereits aus c, des allgemeinen Prinzips).

Da Operationen auch parametrisiert werden können, bedeutet dies nicht, dass vom selben Typ nur endlich viele verschiedene Operationen existieren.

Gegeben eine Menge S von Objekten muss zunächst die Familie $\mathcal{M}$ von Operationenmengen explizit oder implizit spezifiziert werden. In einem praktikablen System kann nicht jede denkbare Operationenmenge betrachtet und ausprobiert werden. Selbst wenn dann $\mathcal{M}$ gegeben ist, kann aus Aufwandsgründen in der Regel nicht sichergestellt werden, dass obiges Optimierungsproblem gelöst werden

kann.

$\Rightarrow$

Bei der Spezifikation von $\mathcal{M}$ und bei der Lösung des daraus resultierenden Optimierungsproblems muss Expertenwissen mit einfließen.

Beispiel 4.1 (Fortführung)

Es macht wenig Sinn, Operationen mit zu berücksichtigen, die nicht zu möglichen Mutationen korrespondieren.

$\Rightarrow$

Hier ist das Expertenwissen des Molekularbiologen gefragt.

Es macht wenig Sinn, die Familie $\mathcal{M}$ derart zu spezifizieren, so dass das resultierende Optimierungsproblem nicht zumindest approximativ zufriedenstellend gelöst werden kann.

$\Rightarrow$

Hier ist der Algorithmiker gefragt. ◇

Beispiel 4.2 (Fortführung)

Eine kompakte Kodierung der Objektmenge könnte z.B. aus einer relativ kleinen Menge von Prototypen, so dass jeder Stuhl aus einem dieser Prototypen mittels Anwendung einiger Operationen aus der Operationenmenge rekonstruierbar ist, bestehen.

Sowohl bei der Spezifikation von $\mathcal{M}$ als auch bei der Lösung des korrespondierenden Optimierungsproblems ist der Computergraphiker gefragt. ◇

Annahme:
Die Spezifikation von $\mathcal{M}$ ist erfolgt und $M \in \mathcal{M}$ ist ausgewählt.

Ziel:
Definition der Kostenfunktion für M.

Man kann die Definition der Kosten einer Operation von der Häufigkeit ihrer Anwendung oder auch von der Kompliziertheit der Operation abhängig machen. Hierbei kann die Kompliziertheit von der Länge des benötigten Codes oder von der benötigten Zeit für die Durchführung einer Operation oder auch von beiden abhängen.

Beispiel 4.1 (Fortführung)

Die Kosten einer Mutation sollten mit der Häufigkeit, mit der sie in der evolutionären Geschichte aufgetreten sind, zu tun haben. Je größer die Wahrscheinlichkeit, dass eine Mutation auftritt, umso geringer sollten die Kosten für diese Mutation sein. D.h., die Definition der Kosten einer Operation sollte in Abhängigkeit der Häufigkeit ihrer Anwendung in

der evolutionären Geschichte erfolgen.

Beispiel 4.2 (Fortführung)

Hier sagt die Intuition, dass einfache Operationen r, deren inverse Operation r^{-1} auch einfach ist, zu ähnlichen Teilen zweier Stühle korrespondieren. Falls mindestens eine der Operationen r und r^{-1} schwierig ist, dann scheinen diese Operationen zu weniger ähnlichen Teilen der Stühle zu korrespondieren.

$\Rightarrow$
Definition der Kosten einer Operation sollte in Abhängigkeit ihrer Kompliziertheit erfolgen.

Falls die Modellierung derart erfolgt ist, dass nicht jeder Stuhl in jeden transformiert werden kann, dann gibt es Paare von Stühlen in S, die die Ähnlichkeit 0 haben. Wenn dies nicht gewollt ist, dann muss dies bei der Modellierung berücksichtigt werden.

Annahme:
Die Operationenmenge M und die Kostenfunktion $c: M \to \mathbb{Q}$ sind gegeben.

Frage:
Wie berechnet man für gegebene Objekte $x, y \in S$ ihre Distanz $d_c(x,y)$?

Definition von $d_c(x,y)$ $\Rightarrow$

Es muss auch hier ein Optimierungsproblem gelöst werden.

Ideen:

- Berücksichtige bereits bei der Definition der Kostenfunktion, dass das Optimierungsproblem zu lösen ist.
- Löse das resultierende Optimierungsproblem nicht exakt, sondern approximativ.
- Möglicherweise ist die kompakte Repräsentation der Objektmenge S unter Verwendung der Operationsmenge M dergestalt, dass diese zur Lösung des Optimierungsproblems herangezogen werden kann.

Bemerkung:

Bei der Entwicklung obiger Theorie für Ähnlichkeit sind wir davon ausgegangen, dass die zugrunde liegende Objektmenge S statisch und ganz im voraus bekannt ist. MDL und MML erlauben das Hinzukommen von neuen Daten. Ebenso könnten wir das Hinzukommen von neuen Objekten zulassen und die Theorie der Ähnlichkeit entsprechend erweitern.