ADS: Algorithmen und Datenstrukturen 2

Teil XIII

Peter F. Stadler & Konstantin Klemm

Bioinformatics Group, Dept. of Computer Science & Interdisciplinary Center for Bioinformatics, **University of Leipzig**

29. Juni 2011

Matroide (Wdh.)

- Sei E eine endliche Menge und $\mathcal{M}$ eine Menge von Teilmengen von E, also $\mathcal{M} \subseteq \mathcal{P}(X)$. Dann heißt $(E, \mathcal{M})$ *Mengensystem*.
- Ein Mengensystem $(E, \mathcal{M})$ heißt *Unabhängigkeitssystem*, wenn $\emptyset \in \mathcal{M}$ und zu jeder Menge in $\mathcal{M}$ auch alle Teilmengen enthalten sind, also gilt:

  Aus $A \in \mathcal{M}$ und $B \subseteq A$ folgt $B \in \mathcal{M}$.
- Ein Unabhängigkeitssystem $(E, \mathcal{M})$ heißt *Matroid*, wenn für alle $A, B \in \mathcal{M}$ die folgende *Austauscheigenschaft* gilt:
  Ist $|A| > |B|$, dann gibt es $x \in A \setminus B$, so dass $B \cup \{x\} \in \mathcal{M}$.

Greedy-Algorithmus (Wdh.)

- Mengensystem $(E, \mathcal{M})$,
- Gewichtsfunktion $w : E \to \mathbb{R}^+$
- Kanonischer Greedy-Algorithmus:
  Initialisierung $X \leftarrow \emptyset$, $R \leftarrow E$
  Solange $R \neq \emptyset$ {
    Wähle $x \in R$ mit $w(x)$ maximal
    $R \leftarrow R \setminus \{x\}$
    Falls $X \cup \{x\} \in \mathcal{M}$: $X \leftarrow X \cup \{x\}$
  }

Optimalität

Lösung $X \in \mathcal{M}$ ist optimal, wenn für alle $A \in \mathcal{M}$

$$\sum_{x \in X} w(x) \geq \sum_{a \in A} w(a)$$

gilt.

Satz (Greedy-Optimalität):
Der kanonische Greedy-Algorithmus liefert auf $(E, \mathcal{M})$ eine optimale Lösung für beliebige Gewichtsfunktion $w : E \to \mathbb{R}$

$$\Updownarrow$$

$(E, \mathcal{M})$ ist ein Matroid.

Beweis (1)

Vorwärtsimplikation: Matroid $\Rightarrow$ Optimalität

Sei $(E, \mathcal{M})$ Matroid und $w : E \to \mathbb{R}^+$ beliebige Gewichtsfunktion, X Greedy-Lösung und A beliebige Lösung, A maximales Element in $\mathcal{M}$.

Zunächst ist $|X| = |A| =: s$, denn alle maximalen Elemente (Basen) in $\mathcal{M}$ haben dieselbe Kardinalität. Also haben wir die Darstellung

$$X = \{x_1, x_2, \ldots, x_s\}$$
$$A = \{a_1, a_2, \ldots, a_s\}$$

mit Reihenfolgen nach absteigendem Gewicht.

Annahme: X ist nicht optimal, und A hat größeres Gesamtgewicht als X. Dann gibt es einen Index k, so daß $w(a_k) > w(x_k)$. Wähle k minimal.

Beweis (2)

Setze $X' = \{x_1, x_2, \ldots, x_{k-1}\}$ und $A' = \{a_1, a_2, \ldots a_k\}$, womit $|X'| = k - 1 < k = |A'|$.

Wegen der Austauscheigenschaft des Matroids gibt es ein $y \in A' \setminus X'$ so dass $Z := X' \cup \{y\} \in \mathcal{M}$.

Per Konstruktion ist $w(y) > w(x_k)$, also hätte der Greedy-Algorithmus bei der aktuellen Teillösung $x_1, x_2, \ldots, x_{k-1}$ und $x_k, y \in R$ nicht ein Element mit maximalem Gewicht betrachtet, im Widerspruch zur Definition des Greedy-Algorithmus. Also muss die Annahme (X nicht optimal) falsch sein. X ist optimale Lösung, q.e.d.

Beweis (3)

Rückwärtsimplikation: Optimalität $\Rightarrow$ Matroid

Wir beweisen die Kontraposition. Sei also $(E, \mathcal{M})$ ein Mengensystem, das kein Matroid ist. Leteres heisst, dass die Austauscheigenschaft oder der Abschluss unter Teilmengenbildung verletzt ist. In jedem Falle finden wir Mengen $A, B \in \mathcal{M}$ mit $|A| < |B|$, so dass für alle $x \in B \setminus A$ gilt: $A \cup \{x\} \notin B$. Setze $m := |A|$.
Nun wählen wir eine Gewichtsfuktion w auf folgende Weise:

$$w(x) = \begin{cases} m+2 & \text{wenn } x \in A \\ m+1 & \text{wenn } x \in B \setminus A \\ 1/|E| & \text{sonst} \end{cases}$$

Beweis (4)

Gewählte Gewichtsfunktion

$$w(x) = \begin{cases} m+2 & \text{wenn } x \in A \\ m+1 & \text{wenn } x \in B \setminus A \\ 1/|E| & \text{sonst} \end{cases}$$

Nachdem der Greedy-Algorithmus alle Elemente aus A in die Lösung aufgenommen hat, können Elemente aus $B \setminus A$ nicht hinzugefügt werden. Für die Greedy-Lösung X gilt also $X \subseteq E \setminus B$, somit

$$\sum_{x \in X} w(x) \leq m(m+2) + \frac{|E \setminus (A \cup B)|}{|E|} < m^2 + 2m + 1$$

Die Lösung B hat das Gesamtgewicht

$$\sum_{x \in B} w(x) = (m+1)(m+1) = m^2 + 2m + 1 \ .$$

Dieses ist echt größer als der von X. Somit ist X nicht optimal, q.e.d.

Erfüllbarkeitsproblem für Boolesche Variable

- n Boolesche Variable: $x_1, x_2, , \ldots, x_n \in \{0, 1\}$ (falsch/wahr)
- 2 Literale zu jeder Variablen: x_i (Variable selbst), $\bar{x}_i$ (negierte Variable)
- Eine Klausel c ist eine Disjunktion (Oder-Verknüpfung) von Literalen, z.B.

$$c = (x_1 \vee \bar{x}_2)$$

- Ein Boolescher Ausdruck in konjuktiver Normalform ist eine Konjunktion (Und-Verknüpfung) Boolescher Klauseln, z.B.

$$(x_1 \vee \bar{x}_2) \wedge (x_2 \vee x_3 \vee x_4) \wedge \ldots$$

- Erfüllbarkeitsproblem: Gibt es eine Belegung der Variablen, so dass der gegebene Boolesche Ausdruck wahr ist?

Verbundwahrscheinlichkeit und Faktorisierung

- Lösungsansatz benutzt Inferenz-Verfahren aus der Statistik.
- Jedem Vektor $x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ wird Wahrscheinlichkeit $p(x)$ zugewiesen

$$p(x) = \begin{cases} 1/Z & \text{falls } x \text{ Lösung ist.} \\ 0 & \text{sonst} \end{cases}$$

- Z = Anzahl Lösungen. Damit ist p korrekt normiert, also $\sum_{x \in \{0,1\}^n} p(x) = 1$.
- Verbundwahrscheinlichkeit kann als Produkt über Klauseln geschrieben werden

$$p(x) = Z^{-1} \prod_{c \in C} c(x)$$

- Ergebnis des Produkts ist 1, wenn für Belegung x alle Klauseln erfüllt sind, 0 sonst.

Marginale

- Marginal $p_i(x_i)$ bzgl. einer Variablen x_i bekommt man durch Summmieren über alle anderen Variablen, z.B.

$$p_1(x_1) = \sum_{x_2 \in \{0,1\}} \sum_{x_3 \in \{0,1\}} \cdots \sum_{x_n \in \{0,1\}} p(x)$$

- Wenn Marginal $p_i(0) > 0$ ist, weiss man, dass es mindestens eine Lösung x mit $x_i = 0$ gibt und kann die Variable x_i eliminieren. Entsprechend für $p_i(1) > 0$.

Beispiel

Betrachte Booleschen Ausdruck

$$(x_1 \vee \bar{x_2}) \wedge (x_2 \vee x_3 \vee x_4)$$

Lösungsmenge ist

$$X = \{0001, 0010, 0011, 1001, 1010, 1011, 1100, 1101, 1110, 1111\} .$$

Anzahl Lösungen $Z = |X| = 10$.

Marginale:

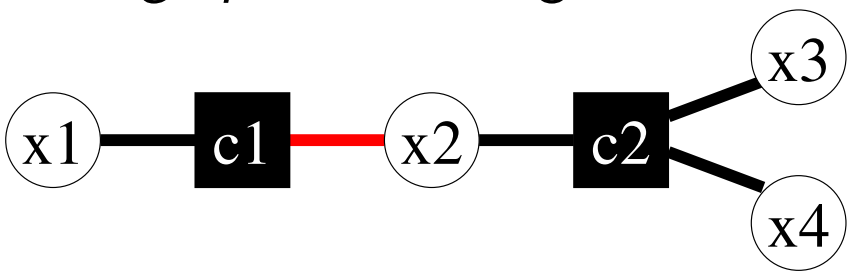
$p_1(0) = 3/10,\ p_1(1) = 7/10$
$p_2(0) = 6/10,\ p_2(1) = 4/10$
$p_3(0) = 4/10,\ p_3(1) = 6/10$
$p_4(0) = 4/10,\ p_3(1) = 6/10$

Faktorgraph

Betrachte Booleschen Ausdruck

$$(x_1 \vee \bar{x_2}) \wedge (x_2 \vee x_3 \vee x_4)$$

Der zugehörige *Faktorgraph* ist ein ungerichteter bipartiter Graph



Zwei Typen von Knoten

- v-Knoten repräsentieren Variable x_i
- f-Knoten repräsentieren Klauseln (Funktionen) c_r
- Kante zwischen v-Knoten i und f-Knoten r existiert, gdw. Klausel c_r von Variable x_i abhängt.

Mitteilungen (message passing)

Versenden von Mitteilungen auf dem Faktorgraphen:
v-Knoten i sendet Mitteilg. $\mu_{i\to r}(x_i)$ an benachbarten f-Knoten r

$$\mu_{i\to r}(x_i) = \prod_{r^*\in N(i)\setminus\{r\}} \nu_{r^*\to i}(x_i)$$

f-Knoten r sendet Mitteilg. $\nu_{r\to i}(x_i)$ an benachbarte v-Knoten i

$$\nu_{r\to i}(x_i) = \sum_{y} c_r(y) \prod_{i^*\in N(r)\setminus\{i\}} \mu_{i^*\to r}(y_i) \ ,$$

wobei $\sum_y$ die Summation über alle Belegungen der Variablen in $N(r)\setminus\{i\}$ läuft.

Aus einer konsistenten Lösung dieser Gleichungen ergibt sich das Marginal $p_i(x_i)$ als

$$p_i(x_i) \propto \prod_{r\in N(r)} \nu_{r\to i} \ .$$

Konvergenz

Ist der Faktorgraph ein Wald, so

- wählt man in jeder Zusammenhangskomponente einen beliebigen Knoten des Faktorgraphen als Wurzel w.
- Von den Blättern beginnend berechnet man iterativ Mitteilungen, die in Richtung w gesendet werden.
- Sind alle Mitteilungen bei w eingetroffen, werden "auslaufende" Mitteilungen (von w wegführend) gesendet, bis bei allen Blättern eine Mitteilung eingetroffen ist.
- Man kann zeigen, dass die so gefundenen Mitteilungen eine konsistente Lösung sind.

Hat der Faktorgraph Zyklen, so

- werden die Mitteilungen so lange neu berechnet, bis sich innerhalb einer Toleranz konstante Werte einstellen (Abbruchkriterium).

Fitness- / Energielandschaften (Wdh.)

- Sei $G = (X, E)$ ein ungerichteter Graph und $f : X \to \mathbb{R}$. Dann heißt

$$(X, E, f)$$

  Wertelandschaft.
- Je nachdem, ob man globale Minima / Maxima sucht, wird (X, E, f) auch Energielandschaft / Fitnesslandschaft genannt.
- Nachbarschaft von Knoten $x \in X$:

$$N(x) := \{y \in X : \{x, y\} \in E\}$$

Metropolis und Simuliertes Abkühlen (Wdh.)

Definiere Markov-Kette auf X durch Übergangswahrscheinlichkeit $P_{x\to y}$ von Knoten (Zustand) $x \in X$ zu $y \in N(x)$

$$P_{x\to y} = \begin{cases} \frac{1}{|N(x)|} & \text{wenn } f(y) \le f(x) \\ \frac{1}{|N(x)|} \exp([f(x) - f(y)]/T) & \text{sonst} \end{cases}$$

und $P_{x\to x} = 1 - \sum_{y\in N(X)} P_{x\to y}$

Simuliertes Abkühlen: Parameter T wird mit der Zeit erniedrigt.

$\to$ Applets zur Illustration.