

Bachelorarbeit

Numerische Analyse des TrueSkill Verfahrens

Johannes Loevenich

Datum der Abgabe

Betreuung: Prof. Dr. Jochen Garcke

Fakultät für Mathematik
Rheinische Friedrich-Wilhelms- Universität Bonn

Inhaltsverzeichnis

1	Einleitung	3
2	Das Ranking Problem	4
3	Bayssche Methodik	4
3.1	Spielausgangsbasierte Updates	5
3.2	Zeitabhängige Updates	5
4	Wahrscheinlichkeitstheorie	5
4.1	Eigenschaften von Zufallsvariablen	8
4.2	Mehrdimensionale Gauß Verteilung	9
5	Faktorgraphen	11
5.1	Allgemeines	11
5.2	Beispiel	11
5.3	Indikatorfunktion und Posterior Wahrscheinlichkeit	12
5.4	Bayssche Netwerke	12
5.5	Sum-Product Algorithmus	12
5.6	Funktionsweise des Algorithmus	13
5.7	Belief Propagation in Baysschen Netzwerken	14
6	Expection Propagation	15
6.1	Assumed-density Filtering	15
7	Ein Ansatz mittels Faktorgraphen für das Trueskill Verfahren	16
7.1	Der Trueskill Faktorgraph	16
7.2	Prior Faktoren	17
7.3	Unsicherheiten	18
7.4	Teamperformanz	18
7.5	Approximieren des Message Passing	19

1 Einleitung

Das Problem n Spieler zu bewerten ist ein wichtiger Forschungsbereich des Maschinellen Lernens.

Aus der Kombinatorik lässt sich schnell herleiten, dass es $n!$ verschiedene Möglichkeiten gibt n Spieler in einem Ranking zu bewerten. Ziel ist es das eine Ranking zu finden, welches das tatsächliche Können der einzelnen Spieler am besten widerspiegelt. Nimmt man an, dass alle Rankings gleich wahrscheinlich sind, so würde dies bedeuten, dass $\log_2(n!) \approx n \log_2(n)$ Spielausgänge nötig wären, um die korrekte Bewertung zu ermitteln. Diese Schranke ist jedoch nur dann aussagekräftig, falls die Ausgänge eines jeden Spiels gleichverteilt sind. Bei vielen Spielen wird durch das matchen von nahezu gleich starken Gegnern versucht diese Chancengleichheit zu erzwingen. In dieser Arbeit wird eine Wahrscheinlichkeitstheoretische Interpretation dieses Problems betrachtet, weshalb es im weiteren Verlauf von Bedeutung ist gewisse Unsicherheiten in Rang eines Spielers zu berücksichtigen. Interessanterweise reduziert sich die minimale Anzahl von aussagekräftigen Spielen auf $n \log_2(m)$, wenn wir annehmen, dass es $m \ll n$ Äquivalenzklassen oder Level gibt.

Betrachten wir Spiele bei denen k Teams gegeneinander antreten und bewertet werden, dann erfüllt jedes Spiel $\log_2(k)$ Ausgänge und es werden nur $\frac{n \log_2(n)}{\log_2(k!)}$ aussagekräftige Spielausgänge benötigt. Man wird sehen, dass das hier vorgestellte Verfahren für das betrachtete Problem nahezu optimal ist, also bis auf geringe Abweichungen gegen die oben genannten Schranken konvergiert.

2 Das Ranking Problem

In dieser Arbeit werden Spiele betrachtet, bei denen mehrere Teilnehmende Teams bewertet werden sollen. Falls zwei Spieler oder Teams dieselbe Bewertung erhalten, sagen wir das das Spiel unentschieden ausgegangen ist. Im Spezialfall von zwei Teams gibt es genau drei mögliche Spielausgänge: Sieg, Niederlage oder unentschieden.

Nummerieren wir alle teilnehmenden Spieler eines Spiels von Eins bis n , so kann ein Spiel zwischen k Teams vollständig durch die k Indizes $i_j \in \{1, \dots, n\}$ der n_j Spieler im j -ten Team beschrieben werden. Der von jedem Team erreichte Rang sei dann definiert als $\mathbf{r} := (r_1, \dots, r_k)^T \in \{1, \dots, k\}^k$. Es wird davon ausgegangen, dass der Gewinner eines Spiels den Rang 1 erhält.

Gesucht wird nach dem Skill s_i eines jeden Spielers, $\mathbf{s} \in \mathbb{R}^n$; wobei der Skill des j -ten Teams eine Funktion $S(s_{i_j})$ in Abhängigkeit aller Skills der Spieler des Teams ist. Im trivialen Fall, dass jedes Team nur einen Spieler enthält beschreibt $S(s_i) = s_i$ also die Identität. Für Skills wird die folgende Eigenschaft gefordert.

Definition 2.1. (Stochastische Transitivität) Falls ein Team u vor einem Team v platziert wurde, so ist es wahrscheinlicher, dass Team u gegen Team v gewinnt als umgekehrt.

$$S(s_{i_u}) \geq S(s_{i_v}) \Rightarrow P(\text{ Team } u \text{ gewinnt }) > P(\text{ Team } v \text{ gewinnt }) \quad (2.1)$$

3 Bayssche Methodik

Natürlich kennt man den Skill eines Spielers nie mit absoluter Sicherheit. Deshalb wird das Problem aus einer Baysschen Sichtweise betrachtet. Dabei wird der Glaube an den Skill eines Spielers durch eine Wahrscheinlichkeitsverteilung $P(\mathbf{s})$ beschrieben. Im folgenden wird angenommen, dass diese Verteilung einer multivariaten Gaußverteilung entspricht, deren Kovarianzmatrix nur Einträge auf der Diagonalen hat ($P(\mathbf{s}) = N(\mathbf{s}, \mu, \text{diag}(\sigma^2))$). Diese Annahme hat einige Vorteile:

1. Die Verteilung ist unimodal in $\mu \in \mathbb{R}^n$. Ein Spieler hat deshalb genau einen unbekannten Skill. Dieser schwankt nicht unerwartet zwischen verschiedenen Spiel Leveln.
2. Die Verteilung impliziert einfache Skillupdate Funktionen. (Siehe **TODO: Kapitel einfügen!**)
3. Die Verteilung kann Speichereffizient implementiert werden. Eine Diagonalmatrix kann leicht als Vektor abgespeichert werden. Der mittlere Skill μ_i und die Skillvarianz σ_i^2 sind für jeden Spieler Konstanten.

3.1 Spielausgangsbasierte Updates

Nachdem der Ausgang eines Spieles in Form eines Vektor, deren Einträge die Platzierungen der Spieler widerspiegeln, wird die Verteilung der Skills aller Spieler $P(s)$ mithilfe der Baysschen Regel aktualisiert:

$$P(s|r, \{i_1, \dots, i_k\}) = \frac{P(r|s, \{i_1, \dots, i_k\})P(s|\{i_1, \dots, i_k\})}{P(r|\{i_1, \dots, i_k\})} = \frac{P(s|S(s, i_1), \dots, S(s, i_k))P(s)}{P(r|\{i_1, \dots, i_k\})} \quad (3.1)$$

Diese neue Verteilung wird *Posterior* Verteilung genannt und wird im darauf folgenden Spiel für $P(s)$ benutzt. Diese Art der Vorgehensweise ist als *on-line learning* bekannt: Jederzeit existiert nur eine Verteilung $P(s)$, die durch jeden Spielausgang beeinflusst wird. Es wird außerdem angenommen, dass die *Posterior* Verteilung alle verfügbaren Informationen enthält. In der Praxis ist es schwer diese Größe exakt, kompakt und effizient darzustellen. Deshalb wird sie meist z.B. mithilfe von Methoden, wie ADF approximiert und mit der approximierten Größe weiter gerechnet. Mit solchen Approximationen wird sich in Kapitel befasst. **TODO: Kapitel einfügen!**

3.2 Zeitabhängige Updates

4 Wahrscheinlichkeitstheorie

Dieses Kapitel gibt einen Überblick über einige wichtige Konzepte der Wahrscheinlichkeitstheorie. Im Folgenden werden Mengen mit Großbuchstaben, z.B. X , und deren Elemente mit Kleinbuchstaben, z.B. x , bezeichnet. Für Mengen definiere die Indikatorfunktion I_X durch

$$I_X(x) := \begin{cases} 0 & \text{falls } x \notin X \\ 1 & \text{falls } x \in X \end{cases}$$

Definition 4.1. (σ -Algebra) Sei eine Menge χ gegeben. Ein Mengensystem \mathcal{Y} von Mengen $X \subseteq \chi$ wird genau dann σ -Algebra über χ genannt, wenn

1. Ist eine Menge $X \in \mathcal{Y}$ enthalten, so auch ihr Komplement $X^c = \chi \setminus X$.
2. Falls $X_i \in \mathcal{Y}$, $i = 1, \dots, \infty$ abzählbares Mengensystem in \mathcal{Y} , dann sind auch $\cup_{i=1}^{\infty} X_i \in \mathcal{Y}$ und $\cap_{i=1}^{\infty} X_i \in \mathcal{Y}$ in \mathcal{Y} enthalten.

Kurz, jede σ -Algebra ist abgeschlossen unter Komplementbildung und abzählbaren Vereinigungen oder Schnitten.

Definition 4.2. (Borelmengen) Sei $\chi = \mathbb{R}^n$, die Borelmengen B_n sind die kleinsten σ -Algebren, die alle offenen Intervalle

$$\{(x_1, \dots, x_n) \in \mathbb{R}^n | \forall i \in \{1, \dots, n\} : x_i \in (a_i, b_i)\}$$

für alle $a_i, b_i \in \mathbb{R}$. Bemerke, dass B_n überabzählbar ist.

Definition 4.3. (Maß- und Wahrscheinlichkeitsraum) Ein messbarer Raum ist ein Tupel (χ, \mathcal{Y}) . Dabei ist χ das Universum und \mathcal{Y} σ -Algebra über χ . Ein Wahrscheinlichkeitsraum ist ein Tripel (χ, \mathcal{Y}, P) , wobei P ein Wahrscheinlichkeitsmaß auf χ ist. D.h. $P : \mathcal{Y} \rightarrow [0, 1]$, sodass $P(\chi) = 1$ und für disjunkte abzählbare Vereinigungen $X_i \in \mathcal{Y}$, $i = 1, \dots, \infty$ gilt

$$P(\cup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} P(X_i).$$

Definition 4.4. (Messbarkeit) Sei (χ, \mathcal{Y}) messbarer Raum. Eine reellwertige Funktion $g : \chi \rightarrow \mathbb{R}$ heißt \mathcal{Y} -messbar (oder messbar) genau dann, wenn

$$\forall z \in \mathbb{R} : \{x \in \chi | g(x) \leq z\} \in \mathcal{Y}.$$

Definition 4.5. (Zufallsvariable) Sei (χ, \mathcal{Y}) messbarer Raum. Eine Zufallsvariable ist eine χ -messbare reellwertige Funktion $f : \chi \rightarrow \mathbb{R}$.

Eine Zufallsvariable $Y = f(X)$ induziert also ein Maß P_Y auf \mathbb{R} , für das die σ -Algebra \mathcal{A} die Intervalle der Form $(-\infty, z] | z \in \mathbb{R}$ enthält. Das Maß P_Y ist vom Maß P_X und f induziert. Das bedeutet

$$\forall Y \in \mathcal{B}_1 : P_Y(Y) := P_X(\{x \in \chi | f(x) \in Y\}).$$

Definition 4.6. (Verteilungsfunktion und Dichte) Für eine Zufallsvariable X heißt die durch

$$F_X(x) := P_X(X \leq x)$$

definierte Funktion $F_X : \mathbb{R} \rightarrow [0, 1]$ Verteilungsfunktion von X . Die Funktion $f_X : \mathbb{R} \rightarrow \mathbb{R}$ wird Dichte genannt, falls

$$\forall z \in \mathbb{R} : F_X(z) = \int_{x \leq z} f_X(x) dx.$$

Für weitere Betrachtungen ist Erwartung einer Zufallsvariablen von essentieller Bedeutung.

Definition 4.7. (Erwartungswert) Sei $f : \chi \rightarrow \mathbb{R}$ messbare Funktion. Der Erwartungswert $E_X[f(X)]$ von f über die Wahrscheinlichkeit von x wird durch

$$E_X[f(X)] := \int_{\mathbb{R}} f(x) dF_X(x)$$

definiert. Der Erwartungswert ist nur dann definiert, wenn $\int_{\mathbb{R}} |f(x)| dF_X(x) < \infty$.

Definition 4.8. (Varianz) Die Varianz $\text{Var}(X)$ einer Zufallsvariable X ist definiert durch

$$\text{Var}(X) := E_X[(X - \mu)^2] = E_X[X^2] - \mu^2,$$

wobei $\mu = E_X[X]$ der Erwartungswert der Zufallsvariable X ist.

Definition 4.9. (Produkttraum) Seien zwei Maßräume $(\mathcal{X}, \mathcal{Y})$ und (φ, Φ) gegeben. Definiere den Produkttraum durch $(\mathcal{X} \times \varphi, \mathcal{Y} \times \Phi)$. Hierbei bezeichnet $\mathcal{Y} \times \Phi$ die kleinste σ -Algebra welche die Mengen $\{X \times Y | X \in \mathcal{Y}, Y \in \Phi\}$ enthält.

Definition 4.10. (Marginal und bedingte Wahrscheinlichkeiten) Sei $(\mathcal{X} \times \varphi, \mathcal{Y} \times \Phi, P_{XY})$ der Produkttraum von X und Y . Das Marginalwahrscheinlichkeitsmaß P_X ist dann durch

$$\forall X \in \mathcal{Y} : P_X(X) := P_{XY}(X \times \varphi)$$

definiert. Sei $Y \in \Phi$ und $P_Y(Y) > 0$, dann ist das bedingte Wahrscheinlichkeitsmaß $P_{X|Y \in \mathcal{Y}}$ durch

$$\forall Y \in \mathcal{Y} : P_{X|Y}(X) := \frac{P_{XY}(X \times Y)}{P_Y(Y)}$$

gegeben.

Definition 4.11. Unabhängigkeit Zwei Zufallsvariablen X und Y werden genau dann unabhängig genannt, wenn

$$\forall X \in \mathcal{Y} : \forall Y \in \Phi : P_{XY}(X \times Y) = P_X(X)P_Y(Y).$$

In diesem Fall genügen die Marginalverteilungen um den gesamten Produkttraum zu definieren.

Im Folgenden schreibe \mathbf{X} , falls \mathbf{X} eine Folge (X_1, \dots, X_n) von Zufallsvariablen definiert. Eine solche Folge kann je nach Kontext entweder als Zeilen- oder als Spaltenvektor interpretiert werden. Ein Element des Universums \mathcal{X}^n wird dann durch ein n -Tupel \mathbf{x} beschrieben. Sei $\mathbf{x} = (x_1, \dots, x_n)$ ein solches n -Tupel, dann verstehe $x \in \mathbf{x}$ als $\exists i \in \{1, \dots, n\} : x_i = x$.

Definition 4.12. (Erwartungswert einer n -dimensionalen Zufallsvariable) Seien $\mathbf{X} = (X_1, \dots, X_n)$ n Zufallsvariablen mit gemeinsamem Wahrscheinlichkeitsmaß P_X , dann ist die Erwartung $E_{\mathbf{X}}[\mathbf{X}]$ durch das n -Tupel

$$E_{\mathbf{X}}[\mathbf{X}] = (E_{X_1}[X_1], \dots, E_{X_n}[X_n])$$

gegeben.

Definition 4.13. (Kovarianz und Konvarianzmatrix) Seien X und Y zwei Zufallsvariablen mit gemeinsamem Wahrscheinlichkeitsmaß P_{XY} , dann ist die Kovarianz $Cov(X, Y)$ durch

$$Cov(X, Y) := E_{XY}[(X - \mu)(Y - \nu)]$$

definiert, wobei $\mu = E_X[X]$ und $\nu = E_Y[Y]$. Es ist leicht einzusehen, dass $Cov(X, X) = Var(X)$ Sei $\mathbf{X} = (X_1, \dots, X_n)$ eine Folge von n Zufallsvariablen und $\mathbf{Y} = (Y_1, \dots, Y_n)$ eine

weitere solche Folge mit gemeinsamer Dichte P_{XY} , dann ist die $n \times m$ Kovarianzmatrix $\text{Cov}(\mathbf{X}, \mathbf{Y})$ definiert durch

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) := \begin{pmatrix} \text{Cov}(X_1, Y_1) & \dots & \text{Cov}(X_1, Y_m) \\ \vdots & \dots & \vdots \\ \text{Cov}(X_n, Y_1) & \dots & \text{Cov}(X_n, Y_m) \end{pmatrix}$$

Falls $\mathbf{X} = \mathbf{Y}$ $\text{Cov}(\mathbf{X}, \mathbf{X}) = \text{Cov}(\mathbf{X})$

4.1 Eigenschaften von Zufallsvariablen

Satz 4.14. (Erwartungswert von Summen und Produkten) Seien X und Y zwei unabhängige Zufallsvariablen, so gilt

$$E_{XY}[X \cdot Y] = E_X[X] \cdot E_Y[Y], \quad (4.1)$$

$$E_{XY}[X + Y] = E_X[X] + E_Y[Y], \quad (4.2)$$

immer dann, wenn die Terme auf der rechten Seite existieren. Die zweite Aussage gilt sogar, falls X und Y nicht unabhängig sind.

Satz 4.15. (Linearität des Erwartungswertes) Für jede n -dimensionale Zufallsvariable \mathbf{X} , jede Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ und jeden stationären Vektor $\mathbf{b} \in \mathbb{R}^m$ gilt

$$E_X[\mathbf{A}\mathbf{X} + \mathbf{b}] = \mathbf{A}E_X[\mathbf{X}] + \mathbf{b}.$$

Satz 4.16. Varianz Zerlegung Seien X und Y zwei unabhängige Zufallsvariablen, dann gilt

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Beweis 1. Setze $\mu = E_X[X]$ und $\nu = E_Y[Y]$. Benutze die Definition der Varianz

$$\begin{aligned} E_{XY}[(X + Y - E_{XY}[X + Y])^2] &= E_{XY}[((X - \mu)(Y - \nu))^2] \\ &= E_{XY}[(X - \mu)^2 + 2 \cdot (X - \mu)(Y - \nu) + (Y - \nu)^2] \\ &= E_X[(X - \mu)^2] + 2E_{XY}[(X - \mu)(Y - \nu)] + E_Y[(Y - \nu)^2] \\ &= E_X[(X - \mu)^2] + 2E_X[(X - \mu)]E_Y[(Y - \nu)] + E_Y[(Y - \nu)^2] \\ &= \text{Var}(X) + \text{Var}(Y) \end{aligned}$$

Lemma 4.17. Für jede Zufallsvariable X und jede Konstant $c \in \mathbb{R}$ gilt $\text{Var}(cX) = c^2 \cdot \text{Var}(X)$.

4.2 Mehrdimensionale Gauß Verteilung

In diesem Abschnitt werden einige der wichtigsten Eigenschaften der Gaussverteilung zusammengefasst.

Definition 4.18. Sei $\mu \in \mathbb{R}^n$ ein Vektor und $A \in \mathbb{R}^{n \times m}$ eine deterministische Matrix. Sei weiter $Y = (Y_1, \dots, Y_m)$ eine Folge von m unabhängigen, normalverteilten Zufallsvariablen Y_i mit Mittelwert Null und Einheitsvarianz ($Y_i \sim \text{Normal}(0, 1)$). Dann wird $X = AY + \mu$ normal- oder gaußverteilt mit Erwartungswert $E_X[X] = \mu$ und Kovarianzmatrix $\text{Cov}(X) = \Sigma = AA'$ genannt. Da das Maß P_X eindeutig durch diese beiden Größen beschrieben wird schreiben wir im Weiteren auch $Y \sim \text{Normal}(\mu, \Sigma)$.

Satz 4.19. Falls $X \sim \text{Normal}(\mu, \Sigma)$, dann besitzt X genau dann eine Dichte f_X , wenn Σ positiv definit ist. Die Dichte f_X ist gegeben durch

$$f_X(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \quad (4.3)$$

Satz 4.20. Sei $X \sim \text{Normal}(\mu, \Sigma)$ eine n -dimensional normalverteilte Zufallsvariable, $A \in \mathbb{R}^{m \times n}$ stationäre Matrix und $b \in \mathbb{R}^m$ stationärer Vektor. Dann ist auch die Zufallsvariable $Y = AX + b$ normalverteilt mit $Y \sim \text{Normal}(A\mu + b, A\Sigma A^T)$.

Satz 4.21. Angenommen $P_{X|Y=y} = \text{Normal}(Xy, \Gamma)$ ist normalverteiltes Wahrscheinlichkeitsmaß, wobei $X \in \mathbb{R}^{m \times n}$ und $\Gamma \in \mathbb{R}^{m \times m}$ stationäre Matrizen für alle Werte $y \in \mathbb{R}^n$ sind. Falls $P_Y = \text{Normal}(\mu, \Sigma)$ normalverteiltes Wahrscheinlichkeitsmaß ist, so gilt

$$P_{Y|X=x} = \text{Normal}(\Psi(X^T \Gamma^{-1} x + \Sigma^{-1} \mu), \Psi). \quad (4.4)$$

$$P_X = \text{Normal}(X\mu, \Gamma + X\Sigma X^T). \quad (4.5)$$

,wobei $\Psi = (X^T \Gamma^{-1} X + \Sigma^{-1})^{-1}$

Beweis 2. Nach Satz () wissen wir, dass

$$f_{Y|X=x(y)} = \frac{f_{X|Y=y} f_Y(y)}{\int_{\mathbb{R}^n} f_{X|Y=y'(x)} f_Y(y') dy'} = \frac{f_{X|Y=y}(x) f_Y(y)}{f_X(x)}.$$

Es fällt auf, dass der Nenner unabhängig von y ist. Betrachte nun deshalb zunächst den Zähler. Mithilfe von Definition ist Letzteres gegeben durch

$$c \cdot \exp\left(-\frac{1}{2}((x - Xy)^T \Gamma^{-1}(x - Xy) + (y - \mu)^T \Sigma^{-1}(y - \mu))\right),$$

wobei $c = (2\pi)^{-\frac{m+n}{2}} |\Gamma|^{-\frac{1}{2}} |\Sigma|^{-\frac{1}{2}}$ unabhängig von x und y ist. Diesen Ausdruck wiederum können wir umschreiben als

$$c \cdot \exp\left(-\frac{1}{2}((y - x)^T C(y - c) + d(x))\right),$$

mit

$$C = X^T \Gamma^{-1} X + \Sigma^{-1},$$

$$Cc = X^T \Gamma^{-1} X + \Sigma^{-1} \mu,$$

$$d(x) = (x - X\mu)^T (\Gamma + X\Sigma X^T)^{-1} (x - X\mu).$$

Da $d(x)$ als Funktion nicht von y abhängt, kann der Term $\exp(-\frac{1}{2}d(x))$ mit in die Konstante c gezogen werden und somit folgt die erste Gleichung des Satzes indem $\Psi := C^{-1}$ gesetzt wird.

Um die Zweite Aussage zu zeigen kann die Definition von $f_X(x)$ benutzt werden, d.h.,

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}^n} c \cdot \exp(-\frac{1}{2}((y' - c)^T C(y' - c) + d(x))) dy' \\ &= c \cdot \exp(-\frac{1}{2}d(x)) \cdot \int_{\mathbb{R}^n} \exp(-\frac{1}{2}((y' - c)^T C(y' - c))) dy' \\ &= c \cdot \exp(-\frac{1}{2}d(x)) \cdot (2\pi)^{\frac{n}{2}} |C|^{-\frac{1}{2}} = c' \cdot \exp(-\frac{1}{2}d(x)) \\ &= c' \cdot \exp(-\frac{1}{2}(x - X\mu)^T (\Gamma + X\Sigma X^T)^{-1} (x - X\mu)), \end{aligned}$$

wobei die dritte Zeile aus der Definition und der Tatsache folgt, dass sich Wahrscheinlichkeitsdichten zu Einer zusammenfassen lassen. Dies zeigt die zweite Aussage.

5 Faktorgraphen

Ein Faktorgraph ist ein bipartiter Graph, der beschreibt wie eine globale Funktion in Abhängigkeit von verschiedenen Variablen und Faktoren durch ein Produkt von lokalen Funktionen ausgedrückt werden kann.

5.1 Allgemeines

In Faktorgraphen wird zwischen zwei Typen von Knoten unterschieden: Die einen Knoten, welche mit Variablen identifiziert werden (nicht ausgefüllte Knoten) und die Anderen, welche mit lokalen Funktionen identifiziert werden (ausgefüllte Knoten). Kanten verbinden Variablenknoten x_i und Funktionenknoten f genau dann, wenn x_i Argument von f ist.

Sei $X = \{x_i\}_{i \in N}$ eine Menge von Variablen bzgl. der Indexmenge $N = \{1, 2, 3, \dots, n\}$. Falls E triviale Teilmenge von N ist, so bezeichne mit X_E die Teilmenge von X , welche durch E induziert wird. Für jedes $i \in N$ nehme die Variable x_i Werte aus dem Alphabet A_i an. Desweiteren wird angenommen, dass A_i für alle $i \in N$ stets endlich ist. Bezeichne eine bestimmte Belegung der Variablen aus X als Konfiguration der Variablen. Diese Konfigurationen können als Kartesisches Produkt $W = \prod_{i \in N} A_i$, den sogenannten Konfigurationsraum verstanden werden.

Ein Element $w = (w_1, \dots, w_n) \in W$, mit $w_i \in A_i$ ist entspricht der Variablenbelegung $x_1 = w_1, \dots, x_n = w_n$.

Im weiteren sind Funktionen mit Urbild W von besonderem Interesse. Sei $g : W \rightarrow R$ eine solche Funktion, auch globale Funktion genannt. Im Moment beziehe sich der Wertebereich auf die Reellen Zahlen. Im Allgemeinen sei jedoch jeder beliebige Semiring erlaubt.

5.2 Beispiel

Beispielsweise jeder lineare Binärblockcode kann durch eine Menge von Paritätstestgleichungen beschrieben werden, sodass jede Gleichung eine Bedingung an das Codewort $x = (x_1, \dots, x_n)$ beschreibt, z.B., $\sum_{i \in E} x_i = 0$. Der zu einem Binärcode mit der angegebenen Paritätsmatrix

$$H = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \quad (5.1)$$

korrespondierende Faktorgraph wird in Figur () gezeigt.

5.3 Indikatorfunktion und Posterior Wahrscheinlichkeit

Anlehnend an das vorherige Beispiel, sei (x_1, \dots, x_n) ein gleichverteilt gezogenes Codewort, dass über einen Speicherlosen Kanal übermittelt werden soll und $y = (y_1, \dots, y_n)$ die gesamte Ausgabe des Kanals. Die gemeinsame *a posteriori* Wahrscheinlichkeit von $\{x_1, \dots, x_n\}$ kann dann durch eine Funktion f ausgedrückt werden

$$f(x_1, \dots, x_n) = \prod_{E \in Q} f_E(x_E) \prod_{i=1}^n f(y_i|x_i). \quad (5.2)$$

Dabei ist für jeden Wert von x_i , $f(y_i|x_i)$ die korrespondierende Likelihood Funktion am entsprechenden Ausgang y_i des Kanals. Die Faktorgraphdarstellung dieses Szenarios ist in Figur (b) dargestellt.

5.4 Bayssche Netwerke

Bayssche Netze sind gerichtete, azyklische graphische Modelle auf einer Menge von Zufallsvariablen. Jeder Knoten v eines Baysschen Netzes wird mit einer Zufallsvariable assoziiert. Bezeichnet $a(v)$, die Menge der Vorgänger des Knotens v , so hat die Wahrscheinlichkeitsverteilung des Baysschen Netzwerkes die Form

$$p(v_1, v_2, \dots, v_n) = \prod_{i=1}^n p(v_i|a(v_i)) \quad (5.3)$$

Falls $a(v_i) = \emptyset$ (d.h. v_i hat keine Vorgänger), so setze $p(v_i|\emptyset) = p(v_i)$. Figur **FIGUR EINFÜGEN** zeigt beispielsweise ein Bayssches Netzwerk, welches die Faktorisierung

$$p(v_1, v_2, v_3, v_4, v_5) = p(v_1|v_2)p(v_2)p(v_3|v_2, v_4)p(v_4)p(v_5|v_4)$$

beschreibt.

5.5 Sum-Product Algorithmus

Sei $g(X)$ globale Funktion über die Variablen der Menge $X = \{x_i : i \in N\}$, wobei Variable x_i Werte der endlichen Menge A_i annimmt. In diesem Abschnitt wird ein Algorithmus beschrieben, um die marginalen Funktionen

$$G_i(x_i) = \sum_{x_1 \in A_1, \dots, x_{i-1} \in A_{i-1}, x_{i+1} \in A_{i+1}, \dots, x_n \in A_n} g(x_1, \dots, x_n) \quad (5.4)$$

für Variablen $x_i, i \in N$ zu berechnen. Im Weiteren sei $\sum_{x_i} f(x_i) = \sum_{x_i \in A_i} f(x_i)$ und genauso sei für eine Teilmenge $J \subset N$ mit $\sum_{x_i; i \in J} f(X)$ die Summe über alle möglichen Konfigurationen der Variablen x_i über J gemeint. Damit gilt $G_i(x_i) = \sum_{x_j; j \in N \setminus \{i\}} g(X)$. Die Definition der marginalen Funktion G kann nun auf eine Teilmenge J von N ausgeweitet werden.

$$G_i(x_i) = \sum_{x_j; j \in N \setminus J} g(X). \quad (5.5)$$

Falls $g(x_1, \dots, x_n)$ eine Wahrscheinlichkeitsverteilung beschreibt, so ist $G_i(x_i)$ die Marginalverteilung und $G_J(X_J)$ die gemeinsame Wahrscheinlichkeitsverteilung der Variablen über die Indexmenge J .

Ist die Anzahl n der Argumente von g klein, so nutze eine alternative Kurzschreibweise für die Marginalfunktionen. Schreibe statt einem Argument x_i von g ein $+$ um anzudeuten, dass über diese Variable summiert wird.

5.6 Funktionsweise des Algorithmus

Der Sum-product Algorithmus operiert mithilfe einer "message passing"Prozedur, die Produkte von lokalen Funktionen entlang der Pfade des Faktorgraphen aufammelt. Es wird angenommen, dass dieser Graph ein Baum ist, d.h. dass dieser Graph keine Kreise enthält. Die Beschreibung des Algorithmus kann durch die Annahme vereinfacht werden, dass jeder Knoten wie ein Prozessor Nachrichten über Kanten übermittelt und empfängt.

Für diese vereinfachte Betrachtung arbeitet der Algorithmus wie folgt. Die Basisoperation an jedem Knoten ermittelt das Produkt aller eingehenden Nachrichten an diesem Knoten. Für Knoten, die eine Menge von Variablen darstellen, wird dieses Produkt um die zugehörigen lokalen Funktionen erweitert. Die so ermittelten Produkte werden dann mit dem Vorbehalt, dass Nachrichten über ausgehende Knoten keine Faktoren enthalten, über die ausgehenden Kanten übermittelt. Da vorausgesetzt wurde, dass der Faktorgraph keine Kreise enthält, enthält das Produkt der ausgehenden Nachrichten einer Kante und der empfangenen Nachrichten dieser Kante alle Faktoren der globalen Funktion.

Diese sogenannte "message-passing"Prozedur wird von den Blättern des Faktorgraphs aus gestartet und iteriert über alle Knoten des Graphen. An Blättern, die Variablenmengen darstellen entspricht die ausgehende Nachricht einer Representation der lokalen Funktion dieses Knotens. An Blättern, die eine Variable darstellen entspricht sie hingegen der Indikatorfunktion. Alle anderen Knoten des Graphen warten zunächst bis sie genügend Nachrichten gesammelt haben um eine ausgehende Nachricht zu produzieren. Genauer soll das heißen, sie warten solange, bis an jeder bis auf einer eingehenden Kante Nachrichten empfangen wurden. Tritt dieser Fall ein, so wird das Produkt aller eingehenden Nachrichten mit der lokalen Funktion gebildet und über die freie Kante übermittelt. Wird an dieser übrig gebliebenen Kante eine Nachricht empfangen, so werden die Produkte der zugehörigen lokalen Funktion über alle anderen ausgehenden Kanten versendet. Diese Prozedur wird in Figur **FIGUR EINFÜGEN!!!** illustriert.

Dieser Algorithmus wird dann effektiv, wenn man beachtet, dass von lokalen Funktionen über Pfade gesammelte Produkte marginalisiert werden können. D.h., dass nicht

alle Variablen entlang einer Kante beachtet werden müssen. Im Allgemeinen muss eine Variable beachtet werden, wenn sie Argument einer nachfolgenden lokalen Funktion ist. Andernfalls kann sie vernachlässigt werden.

Figur **FIGUR EINFÜGEN!!!** zeigt das Fragment eines Faktorgraphen. Die Update Regeln für dieses Fragment ergeben sich dann also

$$\mu_{x \rightarrow A}(x) = \mu_{B \rightarrow x}(x) \cdot \mu_{C \rightarrow x}(x) \quad (5.6)$$

$$\mu_{A \rightarrow x}(x) = \sum_{y,z} f_A(x, y, z) \cdot \mu_{y \rightarrow A}(x) \cdot \mu_{z \rightarrow A}(x) \quad (5.7)$$

$$F_x(x) = \mu_{x \rightarrow A}(x) \cdot \mu_{A \rightarrow x}(x) \quad (5.8)$$

5.7 Belief Propagation in Baysschen Netzwerken

FIGUR EINFÜGEN!!!

Die Verteilungsfunktion () eines Baysschen Netzwerkes erlaubt eine intuitive Umformulierung zur Representation eines Faktorgraphen. Eine zu einem einzigen Faktor gehörende lokale Funktion in (), hat die Form $f(x|a(x))$, wobei $a(x)$ die Menge der Nachfolger von x im zugehörigen Baysschen Netzwerk ist. In Faktorgraphen wurden Nachfolgerknoten durch Pfeile gekennzeichnet. Diese Pfeile erlauben es uns einen Faktorgraphen ebenso als Bayssches Netzwerk ansehen zu können.

6 Expectation Propagation

In diesem Kapitel werden rekursive Approximationstechniken beschrieben, um die KL-Divergenz zwischen Posterior und Approximation zu minimieren. Das sogenannte Assumed-density Filtering ist eine schnelle Methode auf diesem Gebiet. Expectation Propagation oder kurz EP ist eine Erweiterung des Assumed-density Filtering um Situationen stapelweise zu verarbeiten. Es hat höhere Genauigkeit als das Assumed-density Filtering und andere vergleichbare Methoden um Schlussfolgerungen zu approximieren.

6.1 Assumed-density Filtering

Dieser Abschnitt fasst die Idee des Assumed-density Filtering (ADF) zusammen um die Grundlagen für die Methode des Expectation Propagation zu schaffen. Assumed-density Filtering ist eine der grundlegenden Techniken, um Posterior in Baysschen Netzwerken und anderen statistischen Modellen zu approximieren.

7 Ein Ansatz mittels Faktorgraphen für das Trueskill Verfahren

Das Trueskill Verfahren ist ein Bayssches Ranking Verfahren, dass in seinen Grundzügen eine Erweiterung des Elo Ranking Systems aus dem Schach ist. Dieses neue System beachtet Unsicherheiten in der Bewertung eines Spielers, stellt neue Modelle für ausgeglichene Spiele auf und kann die Skills eines einzelnen Spielers aus dem Rang eines Teams ermitteln. Unsicherheiten werden durch eine Approximierung des message passing in faktorgraphen simuliert.

7.1 Der Trueskill Faktorgraph

Sei $1, \dots, n$ eine Menge von Spielern, die in k Teams gegeneinander antreten. Die Zuweisung der Teams sei durch k disjunkte Teilmengen $A_j \subset \{1, \dots, n\}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$ eindeutig beschrieben. Der Ausgang des Spiels sei definiert durch $\mathbf{r} := (r_1, \dots, r_k) \in \{1, \dots, k\}$, wobei r_j der Rang eines jeden Teams j sei. Falls $r_j = 1$, so hat Team j das Spiel gewonnen. Falls $r_i = r_j$ für $i \neq j$ so sagen wir, dass Team i gegen Team j unentschieden gespielt hat.

Wir wollen die Wahrscheinlichkeit $P(r|\mathbf{s}, A)$ eines bestimmten Spielausganges \mathbf{r} mit gegebenen Skills \mathbf{s} der Spieler und einer Teamzuweisung $A := (A_1, \dots, A_k)$ modellieren. Mithilfe der Bayschen Regel erhalten wir für die Posterior Verteilung

$$p(\mathbf{s}|\mathbf{r}, A) = \frac{P(\mathbf{r}|\mathbf{s}, A)p(\mathbf{s})}{P(\mathbf{r}|A)}. \quad (7.1)$$

Dabei nehmen wir an, dass $p(\mathbf{s}) := \prod_{i=1}^n \mathcal{N}(s_i; \mu_i, \sigma_i^2)$. Dies macht Sinn, da der Faktorgraph selbst eine Menge von Multiplikationen darstellt, um eine gemeinsame Verteilung zu finden. Wir weisen außerdem jedem Spieler in einem Spiel eine Performanz $p_i \sim \mathcal{N}(p_i; s_i, \beta^2)$ mit Zentrum s_i und Varianz β^2 zu. Die Performanz t_j eines Teams j entspricht der Summe $t_j := \sum_{i \in A_j} p_i$ der Spieler des Teams.

Ordnen wir die Teams nach ihren Rängen im Spielausgang \mathbf{r} und nehmen wir an, dass unentschieden zunächst nicht zugelassen sind, so erhalten wir

$$P(\mathbf{r}|\{t_1, \dots, t_k\}) = P(\mathbf{r}|\{t_r(1), \dots, t_r(k)\}). \quad (7.2)$$

Sind unentschieden zugelassen, so erfordert der Spielausgang $r(j) < r(j+1)$ $t_r(1) > t_r(j+1) + \epsilon$ und ein unentschieden der Art $r(j) = r(j+1)$ umgekehrt $|t_r(1) - t_r(j+1)| \leq \epsilon$, wobei $\epsilon > 0$ ein Maß für die Gewinnspanne ist.

Wir wollen nun nach jedem Spiel die Skills der Spieler in Abhängigkeit des Spielausganges \mathbf{r} aktualisieren und nutzen dafür die Darstellung mittels Faktorgraphen und den daran gebundenen Expectation Propagation Algorithmus. Dazu approximieren wir die Posterior Verteilungen gaußverteilt und verwenden sie im darauf folgenden Spiel als Prior Verteilungen.

Um das Verfahren an einem Beispielgraphen herleiten und erklären zu können, betrachten wir folgendes Spiel. Wir nehmen an, dass $k = 3$ Teams mit $A_1 = \{1\}, A_2 =$

$\{2, 3\}$ und $A_3 = \{4\}$ gegeneinander antreten. Der Spielausgang sei $\mathbf{r} = (1, 2, 2)$. Dies bedeutet, dass Team 1 das Spiel gewonnen hat und Team 2 und Team 3 unentschieden gespielt haben. In Figur () geben wir den korrespondierenden Trueskill Faktor Graphen an.

Wie bereits in Kapitel () beschrieben, ist die grundlegende Idee eine gemeinsame Wahrscheinlichkeitsverteilung von mehreren Zufallsvariablen in zwei verschiedene Typen von Knoten aufzuteilen. Schwarze Boxen stellen Faktorknoten und Kreise Variablenknoten dar. Die gemeinsame Verteilung ergibt sich dann als Produkt der einzelnen Faktoren:

$$p(X) = \frac{1}{Z} \prod_S f_S(X_S) \quad (7.3)$$

X_S beschreibt die zu einem Faktor f_S zugehörigen Variablen und Z sei Normalisierungskonstante. Das wichtigste Konzept ist dann die Anwendung der Summen-Produkt-Update Regel:

Jede Nachricht, die von einem Knoten v über eine Kante e übermittelt wird, entspricht dem Produkt der lokalen Funktion an v (Oder der Einheitsfunktion, falls v Variablenknoten ist) mit allen eingegangenen Nachrichten an v außer an Kante e .

In Anlehnung an Kapitel () ergeben sich damit die drei folgenden Gleichungen

$$p(v_k) = \prod_{f \in F_{v_k}} m_{f \rightarrow v_k}(v_k) \quad (7.4)$$

Gleichung () sagt uns, dass der Wert des Marginales an v_k dem Produkt der eingehenden Nachrichten entspricht.

$$m_{f \rightarrow v_j}(v_j) = \int \dots \int f(v) \prod_{i \neq j} m_{v_i \rightarrow f}(v_i) d\mathbf{v}_{\setminus j} \quad (7.5)$$

Gleichung () zeigt, dass der Wert einer Nachricht von einem Faktor f zu einer Variable v_j der Summe über das Produkt aller anderen Nachrichten zwischen diesem Faktor und abhängigen Variablen außer v_j entspricht.

$$m_{v_k \rightarrow f}(v_k) = \prod_{f' \in F_{v_k} \setminus \{f\}} m_{f' \rightarrow v_k}(v_k) \quad (7.6)$$

Gleichung () beschreibt, dass die Nachricht von einer Variable v_k zu einem Faktor f das Produkt aller anderen Nachrichten ist.

Für weitere Details möchten wir auf Kapitel () und die dortigen Beispiele verweisen.

7.2 Prior Faktoren

Betrachten wir die erste Ebene des Trueskill Graphen Prior Faktor ist im Trueskill Graphen als schwarze Box dargestellt.

Um die Trueskill Update Formeln vereinfacht darstellen zu können, verwenden wir die folgenden Notationen

$$\begin{aligned}\pi &= \sigma^{-2} = \frac{1}{\sigma^2} \\ \tau &= \pi\mu = \frac{\mu}{\sigma^2}\end{aligned}\tag{7.7}$$

Das Ziel des Prior Faktors ist es diese Werte einer gaußverteilten Zufallsvariable mit Erwartung m und Varianz $\nu\nu^2$ zu aktualisieren. Mit obiger Notation ergibt sich also:

$$\begin{aligned}\pi_x^{\text{new}} &\leftarrow \pi_x + \frac{1}{\nu^2} \\ \tau_x^{\text{new}} &\leftarrow \tau_x + \frac{m}{\nu^2}\end{aligned}\tag{7.8}$$

Wir sehen, dass die beiden Gleichungen den ersten Gleichungen im Trueskill Paper () entsprechen.

7.3 Unsicherheiten

Wie bereits in der Einleitung erwähnt können Skills nur mit einer gewissen Unsicherheit beschrieben werden. Wir beschreiben hier, wie solche Unsicherheiten im Trueskill Verfahren berücksichtigt werden. Dazu nehmen wir an, dass wir eine gaußverteilte Zufallsvariable y gegeben haben und eine neue gaußverteilte Zufallsvariable $\mathcal{N}(x; y, c^2)$ erhalten wollen, die eine gewisse Unsicherheit c^2 berücksichtigt. Dazu sei $a := (1 + c^2(\pi_y - \pi_{f \rightarrow y}))^{-1}$. Durch Umformungen erhalten wir

$$\begin{aligned}a &= (1 + c^2(\pi_y - \pi_{f \rightarrow y}))^{-1} = \frac{1}{1 + c^2(\pi_y - \pi_{f \rightarrow y})} \\ &\approx \frac{\sigma_y^2}{\sigma_y^2 + c^2}\end{aligned}\tag{7.9}$$

Es ist somit leicht zu sehen, dass a stets kleiner als 1 ist. Nun kann die Ungewissheit c^2 approximativ den Werten π und τ zugewiesen werden:

$$\begin{aligned}\pi_{f \rightarrow x}^{\text{new}} &\leftarrow a(\pi_y - \pi_{f \rightarrow y}) = \frac{1}{\sigma_y^2 + c^2} \\ \tau_{f \rightarrow x}^{\text{new}} &\leftarrow a(\tau_y - \tau_{f \rightarrow y}) = \frac{\mu_y}{\sigma_y^2 + c^2}\end{aligned}\tag{7.10}$$

7.4 Teamperformanz

Dieser Faktor wird benutzt um die Differenzen zwischen Teams innerhalb eines Spieles auszudrücken und summiert mehrere gaußverteilte Variablen miteinander. Wir nehmen

an, dass wir n Variablen v_1, \dots, v_n miteinander summieren wollen, wobei jede dieser Variablen mit einem Faktor a_n gewichtet wird und v_0 dem Wert der Summe entspricht. Es gilt also:

$$v_0 = a_1 v_1 + \dots + a_n v_n \quad (7.11)$$

Aus Kapitel () über die Gaußverteilung wissen wir, dass dies einer Zufallsvariable mit Erwartungswert und Varianz der Form

$$\begin{aligned} \mu &= \sum_{i=1}^n a_i \mu_i \\ \sigma^2 &= \sum_{i=1}^n a_i^2 \sigma_i^2 \end{aligned} \quad (7.12)$$

entpricht. Da die Update Formeln durch den Gebrauch der oben eingeführten Notation mithilfe von π und τ einfacher und kompakter zu implementieren sind, müssen diese in eine solche Form gebracht werden.

7.5 Approximieren des Message Passing

Literatur

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen, als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet habe.

Ort, den Datum