Lernen einer Genomrepräsentation mit einer beschränken Boltzmann-Maschine

Hintergrund

Whole-Genome-Regression

Der Genotyp eines Individuums sei gegeben durch einen Vektor $\mathbf{x} \in \{0,1\}^n$. Dabei sind die x_i die Ausprägungen der Single Nucleotide Polymorphisms (SNPs), die bei einer Person gemessen wurden, etwa durch Sequenzierung oder ein Microarray. Für die Zwecke dieser Aufgabe nehmen wir an, dass die Eingabevariablen binär sind, was etwa dadurch erreicht werden kann, dass die drei Genotypen (AA, Aa, aa) an einer Site jeweils mit zwei Komponenten von $\mathbf{x}$ kodiert sind. Wir betrachten das Problem der Whole-Genome-Regression, bei dem wir einen Phänotyp y, z.B. die Körpergröße des Individuums, nur unter Verwendung von $\mathbf{x}$ vorhersagen wollen. Ein verbreiteteter, einfacher Ansatz ist die lineare Regression

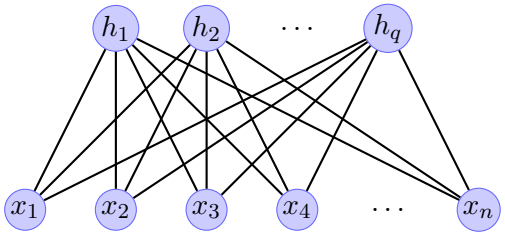
$$\hat{y} = \theta_0 + \sum_i \theta_i x_i. \tag{1}$$

Da n typischerweise groß ist, ist Überanpassung (overfitting) ein Problem, das es in einem solchen Ansatz durch Regularisierung zu vermeiden gilt, etwa durch "Bestrafung" von betragsmäßig großen θ_i mit einem Regularisierungsterm der Form $\sum_i |\theta_i|^r$ oder von vielen $\theta_i \neq 0$. Zur effektiven Regularisierung ist eine gute Modellierung von $P(\mathbf{x})$ hilfreich, die insbesondere die in einer Population gegebenen Abhängigkeiten zwischen verschiedenen SNPS x_i und x_j berücksichtigt. Dazu zählen insbesondere die durch Linkage Disequilibrium (LD) verursachte positive Korrelation von x_i, die nahe im Genom gelegene SNPs repräsentieren, aber auch solche Abhängigkeiten, die durch die Populationsstruktur verursacht ist.
Je nach betrachtetem Phänotyp y kann (1) auch deswegen schlecht geeignet sein, da Interaktionen zwischen verschiedenen x_i bei ihrem Einfluß auf y bestehen können (z.B. Epistasis mit $y = \theta_1 x_1 + \theta_2 x_2 + \theta_{i,j} x_1 x_2$ mit $\theta_{i,j} \neq 0$).

Beschränkte Boltzmann-Maschinen

Restricted Boltzmann Machines (RBMs) sind spezielle graphische Modelle (paarweises Markow'sches Zufallsfeld binärer Variablen über einem bipartiten Graphen), die mit großem Erfolg für das sogenanne unüberwachte *Pretraining* von künstlichen neuronalen Netzen verwendet werden (Stichwort "*deep learning*"). Mit ihnen können automatisch Merkmale in einer Trainingsmenge von ungelabelten Eingaben $\mathbf{x}$ trainiert werden, etwa kleinere Striche und Stiftschwünge bei dem Problem der Handschrifterkennung.

Seien $n, q \in \mathbb{N}$ und $X_1, X_2, \ldots, X_n, H_1, H_2, \ldots, H_q \in \{0,1\}$ binäre Zufallsvariablen (hier *sichtbare* und *versteckte* Einheiten genannt), so dass



$$P(\mathbf{x}, \mathbf{h}) = P(\mathbf{X} = \mathbf{x}, \mathbf{H} = \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{x},\mathbf{h})},$$

wobei

$$\begin{aligned} -E(\mathbf{x}, \mathbf{h}) &= \mathbf{h}^T W \mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h} \\ &= \sum_{i,j} h_i w_{i,j} x_j + \sum_j c_j x_j + \sum_i b_i h_i. \end{aligned}$$

$W = (w_{i,j})$ ist eine $q \times n$-Matrix, $\mathbf{c}$, $\mathbf{b}$ sind Spaltenvektoren mit n bzw. q Komponenten. E wird *Energie* genannt und die Konstante $Z = Z(W, \mathbf{c}, \mathbf{h}) > 0$ ist so gewählt, dass P eine

Wahrscheinlichkeitsverteilung ist:

$$Z = \sum_{\mathbf{x}' \in \{0,1\}^n} \sum_{\mathbf{h}' \in \{0,1\}^q} e^{-E(\mathbf{x}', \mathbf{h}')}$$

Fragestellungen

- Verwende RBMs, um die Genotypen in einer gegebenen Population zu repräsentieren.
- Visualisiere die Gewichte des trainierten RBMs. Sind dort Abhängigkeiten repräsentiert, die nicht durch LD erklärt werden können?
- Können nach unüberwachtem Pretraining bessere Genauigkeiten bei der Vorhersage erzielt werden?
- Wie verändern sich die Antworten auf obige Fragen, wenn beim Training des RBMs die y-Werte mit eingegeben werden (überwachtes Lernen)?

Eingabedaten

Zum Trainieren des Modells und zum Testen stehen die Genotyp-Daten, Körpergrößen und Kovariablen (Alter, Geschlecht) von ca. 4000 Individuen zur Verfügung. Zum Testen von Eigenschaften des Modells können Daten auch simuliert werden.

Implementation

Es kommt in Frage, die RBM und das Training selbst zu implementieren oder auch die Adaption einer bestehenden Bibliothek wie die Shark Open-Source `C++`-Bibliothek zum Maschinellen Lernen.

Literatur

- Asja Fischer und Christian Igel, "Training restricted Boltzmann machines: An introduction", *Pattern Recognition*, 2014
- Ronald de Vlaming und Patrick J. F. Groenen, "The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics", *BioMed Research International*, 2015
- Folienskript von M. Stanke zu "Maschinelles Lernen", `http://bioinf.uni-greifswald.de/bioinf/teaching/ml16/`