Boltzmannův stroj

Organizační dynamika:

- Cyklická síť se symetrickými spoji (tj. libovolný neorientovaný graf)
- Množinu všech neuronů značíme N
- označme ξ_j vnitřní potenciál a y_j výstup (stav) neuronu j
- stav stroje: $\vec{y} \in \{-1, 1\}^{|N|}$.
- označme w_{ji} *reálnou* váhu spoje od neuronu i k neuronu j.
- žádný neuron nemá bias a přepokládáme $w_{jj} = 0$ pro $j \in N$.

Botzmannův stroj

Aktivní dynamika: Stavy neuronů jsou iniciálně nastaveny na hodnoty z množiny $\{-1, 1\}$, tj. $y_j^{(0)} \in \{-1, 1\}$ pro $j \in N$.

V t-tém kroku aktualizujeme náhodně vybraný neuron $j \in N$ takto: nejprve vypočteme vnitřní potenciál

$$\xi_j^{(t-1)} = \sum_{i \in j_{\leftarrow}}^{n} w_{ji} y_i^{(t-1)}$$

a poté náhodně zvolíme hodnotu $y_j^{(t)} \in \{-1, 1\}$ tak, že $\mathbf{P}\left[y_j^{(t)} = 1\right] = \sigma(\xi_j^{(t-1)})$ kde

$$\sigma(\xi) = \frac{1}{1 + e^{-2\xi/T(t)}}$$

Parametr $T(t)$ se nazývá **teplota** v čase t.

Boltzmannův stroj

- Velmi vysoká teplota $T(t)$ znamená, že $\mathbf{P}\left[y_j^{(t)} = 1\right] \approx \frac{1}{2}$ a stroj se chová téměř náhodně.
- Velmi nízká teplota $T(t)$ znamená, že buď $\mathbf{P}\left[y_j^{(t)} = 1\right] \approx 1$ nebo $\mathbf{P}\left[y_j^{(t)} = 1\right] \approx 0$ v závislosti na tom, jestli $\xi_j^{(t)} > 0$ nebo $\xi_j^{(t)} < 0$. Potom se stroj chová téměř deterministicky (tj. jako Hopfieldova síť).

Boltzmannův stroj - reprezentace rozložení

Cíl: Chceme sestrojit síť, která bude reprezentovat dané pravděpodobnostní rozložení na množině vektorů $\{-1, 1\}^{|N|}$.

Velmi hrubá a nepřesná idea: Boltzmannův stroj mění náhodně svůj stav z množiny $\{-1, 1\}^{|N|}$.

Když necháme B. stroj běžet dost dlouho s fixní teplotou, potom budou frekvence návštěv jednotlivých stavů nezávislé na iniciálním stavu.

Tyto frekvence budeme považovat za pravděpodobnostní rozložení na $\{-1, 1\}^{|N|}$ reprezentované B. strojem.

V adaptivním režimu bude zadáno nějaké rozložení na stavech a cílem bude nalézt konfiguraci takovou, že rozložení reprezentované strojem bude odpovídat zadanému rozložení.

Rovnovážný stav

Fixujme teplotu T (tj. $T(t) = T$ pro $t = 1, 2, \ldots$).

Boltzmannův stroj se po jisté době dostane do tzv. *termální rovnováhy*. Tj. existuje čas t' takový, že pro libovolný stav stroje $\gamma^* \in \{-1, 1\}^{|N|}$ a libovolné $t^* \geq t'$ platí, že

$$p_N(\gamma^*) := \mathbf{P}\left[\vec{y}^{(t^*)} = \gamma^*\right]$$

splňuje $p_N(\gamma^*) \approx \frac{1}{Z} e^{-E(\gamma^*)/T}$ kde

$$Z = \sum_{\gamma \in \{-1,1\}^{|N|}} e^{-E(\gamma)/T} \qquad E(\gamma) = -\frac{1}{2} \sum_{i,j} w_{ij} y_i^\gamma y_j^\gamma$$

tj. Boltzmannovo rozložení

Pozn.: Teorie Markovových řetězců říká, že $\mathbf{P}\left[\vec{y}^{(t^*)} = \gamma^*\right]$ je také dlouhodobá frekvence návštěv stavu γ^*.

Toto platí *bez ohledu na iniciální nastavení neuronů!* Síť tedy reprezentuje rozložení p_N.

Boltzmannův stroj - učení

Problém: Tak jak jsme si jej definovali má Boltzmannův stroj omezenou schopnost reprezentovat daná rozložení.

Proto množinu neuronů N disjunktně rozdělíme na

- množinu **viditelných** neuronů V
- množinu **skrytých** neuronů S.

Pro daný stav viditelných neuronů $\alpha \in \{-1, 1\}^{|V|}$ označme

$$p_V(\alpha) = \sum_{\beta \in \{-1,1\}^{|S|}} p_N(\alpha, \beta)$$

pravděpodobnost stavu viditelných neuronů α v termálním ekvilibriu bez ohledu na stav skrytých neuronů.

Cílem bude adaptovat síť podle daného rozložení na $\{-1, 1\}^{|V|}$.

Boltzmannův stroj - učení

Adaptivní dynamika:
Nechť p_d je pravděpodobnostní rozložení na množině stavů viditelných neuronů, tj. na $\{-1, 1\}^{|V|}$.

Cílem je nalézt konfiguraci sítě W takovou, že p_V odpovídá p_d.

Vhodnou mírou rozdílu mezi rozděleními p_V a p_d je relativní entropie zvážená pravděpodobnostmi vzorů
(tzv. Kullback-Leibler divergence)

$$\mathcal{E}(W) = \sum_{\alpha \in \{-1,1\}^{|V|}} p_d(\alpha) \ln \frac{p_d(\alpha)}{p_V(\alpha)}$$

Boltzmannův stroj - učení

$\mathcal{E}(\vec{w})$ budeme minimalizovat pomocí gradientního sestupu, tj. budeme počítat poslounost matic vah $W^{(0)}, W^{(1)}, \ldots$

- váhy v $W^{(0)}$ jsou inicializovány náhodně blízko 0
- v ℓ-tém kroku (zde $\ell = 1, 2, \ldots$) je $W^{(\ell)}$ vypočteno takto:

$$W_{ji}^{(\ell)} = W_{ji}^{(\ell-1)} + \Delta W_{ji}^{(\ell)}$$

kde

$$\Delta W_{ji}^{(\ell)} = -\varepsilon(\ell) \cdot \frac{\partial \mathcal{E}}{\partial w_{ji}}(W^{(\ell-1)})$$

je změna váhy w_{ji} v ℓ-tém kroku a $0 < \varepsilon(\ell) \leq 1$ je rychlost učení v ℓ-tém kroku.

Zbývá spočítat (odhadnout) $\frac{\partial \mathcal{E}}{\partial w_{ji}}(W)$.

Boltzmannův stroj - učení

Formálním derivováním funkce $\mathcal{E}$ lze ukázat, že

$$\frac{\partial \mathcal{E}}{\partial w_{ji}} = -\frac{1}{T}\left(\left\langle y_j^{(t^*)} y_i^{(t^*)}\right\rangle_{fixed} - \left\langle y_j^{(t^*)} y_i^{(t^*)}\right\rangle_{free}\right)$$

- $\left\langle y_j^{(t^*)} y_i^{(t^*)}\right\rangle_{fixed}$ je průměrná hodnota $y_j^{(t^*)} y_i^{(t^*)}$ v termální rovnováze za předpokladu, že hodnoty viditelných neuronů jsou **fixovány** na počátku výpočtu dle rozložení p_d.
- $\left\langle y_j^{(t^*)} y_i^{(t^*)}\right\rangle_{free}$ je průměrná hodnota $y_j^{(t^*)} y_i^{(t^*)}$ v termální rovnováze bez fixace viditelných neuronů.

Celkově

$$\begin{aligned}\Delta w_{ji}^{(\ell)} &= -\varepsilon(\ell) \cdot \frac{\partial \mathcal{E}}{\partial w_{ji}}(W^{(\ell-1)}) \\ &= \frac{\varepsilon(\ell)}{T}\left(\left\langle y_j^{(t^*)} y_i^{(t^*)}\right\rangle_{fixed} - \left\langle y_j^{(t^*)} y_i^{(t^*)}\right\rangle_{free}\right)\end{aligned}$$

Boltzmannův stroj - učení

Pro výpočet $\left\langle y_j^{(t^*)} y_i^{(t^*)} \right\rangle_{fixed}$ proveď následující:

- Polož $\mathcal{Y} := 0$ a proveď následující akce q krát:
  1. fixuj náhodně hodnoty viditelných neuronů dle rozložení p_d (tj. v průběhu následujících kroků je neaktualizuj)
  2. simuluj stroj po t^* kroků
  3. přičti aktuální hodnotu $y_j^{(t^*)} y_i^{(t^*)}$ k proměnné $\mathcal{Y}$.
- $\mathcal{Y}/q$ bude dobrým odhadem $\left\langle y_j^{(t^*)} y_i^{(t^*)} \right\rangle_{fixed}$ za předpokladu, že q je dostatečně velké číslo.

$\left\langle y_j^{(t^*)} y_i^{(t^*)} \right\rangle_{free}$ se odhadne podobně, pouze se nefixují viditelné neurony (tj. v kroku 1. se zvolí libovolný startovní stav a v následném výpočtu se mohou aktualizovat všechny neurony).

Boltzmannův stroj - učení

Pro upřesnění analytická verze:

$$\left\langle y_i^{(t^*)} y_j^{(t^*)} \right\rangle_{fixed} =$$

$$= \sum_{\alpha \in \{-1,1\}^{|V|}} p_d(\alpha) \sum_{\beta \in \{-1,1\}^{|S|}} \frac{p_N(\alpha,\beta)}{p_V(\alpha)} y_j^{\alpha\beta} y_i^{\alpha\beta}$$

kde $y_j^{\alpha\beta}$ je výstup neuronu j ve stavu (α,β).

$$\left\langle y_i^{(t^*)} y_j^{(t^*)} \right\rangle_{free} = \sum_{\gamma \in \{-1,1\}^{|N|}} p_N(\gamma) y_j^{\gamma} y_i^{\gamma}$$