Zastosowanie metryki Wassersteina w problemie uczenia ograniczonych maszyn Boltzmanna

dr inż. Maksymilian Bujok

Zakład Algebry i Kombinatoryki
Wydział Matematyki i Nauk Informacyjnych

27 marca 2019

Plan prezentacji

- Ograniczone maszyny Boltzmanna
- Problem transportu optymalnego
- Odległość Wassersteina
- Uczenie maszyn Boltzmanna z zastosowaniem odległości Wassersteina
- Podsumowanie

Ograniczone maszyny Boltzmanna

Maszyny Boltzmanna

- Maszyny Boltzmanna możemy rozpatrywać jako sieć losowych węzłów (wierzchołków).
- Każdy węzeł (wierzchołek) możemy traktować jak dwustanowy neuron.
- Do opisu całego układu możemy posługiwać się pojęciem energii układu.
- Stan każdego neuronu zmienia się w czasie o kroku dyskretnym w sposób losowy.
- Układ dąży do stanu o niższej energii, który możemy utożsamiać z optimum globalnym.

Maszyny Boltzmanna

- Rozważmy sieć złożoną z N połączonych parami wierzchołków (neuronów) i $I \subseteq \binom{N}{2}$.
- Każdy z wierzchołków $i \in N$ zwany dalej *neuronem* może przyjmować stany $i \in \{0, 1\}$.
- Cała sieć możemy opisać przy pomocy wektora $x = (x_i)_{i \in N} \in \{0, 1\}^N$.
- Każdy neuron ma funkcję $\theta_i \in \mathbb{R}$.
- Każda para neuronów jest połączona krawędzią o wadze $\theta_{\{i,j\}} \in \mathbb{R}$.

Maszyny Boltzmana

- Dla danych wartości $\theta_i \in \mathbb{R}$ oraz $\theta_{\{i,j\}} \in \mathbb{R}$ energia układu, którą możemy utożsamiać z funkcją oceny, przyjmuje postać:

$$\mathbf{E}(x, \theta) = -\sum_{i \in N} \theta_i x_j - \sum_{\{i,j\} \in I} \theta_{\{i,j\}} x_i x_j, \; x \in \{0,1\}^N. \quad (1)$$

- Maszyna Boltzmanna uaktualnia swój stan w dyskretnym czasie.
- Ten proces nazywany jest próbkowaniem Gibbsa (ang. Gibbs sampling).
- Przejście ze stanu $x^{(t)} \in \{0,1\}^N$ w czasie t to stanu w czasie $t+1$, $x^{(t+1)}$ odbywa się poprzez wybór jednego z neuronów $i \in N$ i jego przejściem do stanu $x_i^{(t+1)} = 1$ z prawdopodobieństwem:

$$Pr(x_i^{(t+1)} = 1 || x^{(t)}) = \sigma(\sum_{\{i,j \in I\}} \theta_{\{i,j\}} x_j^{(t)} + \theta_i), \quad (2)$$

lub $x_i^t = 0$ z prawdopodobieństwem $1 - Pr(x_i^{(t+1)})$, a σ jest funkcją logistyczną.

- Prawdopodobieństwo znalezienia się układu, Maszyny Boltzmanna, w określonym stanie opisuje rozkład Gibbsa-Boltzmanna

$$p(x;\theta) = \frac{\exp\left(-E(x;\theta)\right)}{Z(\theta)},\ x \in \{0,1\}^N, \tag{3}$$

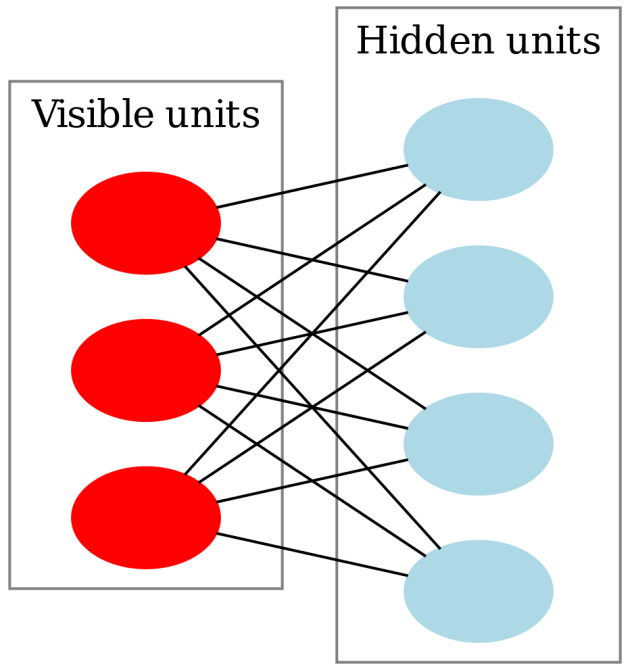
gdzie $Z(\theta) = \sum_x \exp\left(-E(x;\theta)\right.$ nazywamy funkcją rozkładu (sumą statystyczną).

Ograniczone maszyny Boltzmanna

- Będą nas interesowały tzw. **ograniczone maszyny Boltzmanna** (ang. Restricted Boltzmann Machines).
- Zacznijmy od tego, że jest możliwa obserwacja tylko $V \subseteq N$ spośród wszystkich neuronów.
- Pozostałe neurony, czyli $H = N \setminus V$ nazywamy ukrytymi.
- Z rozkładu prawdopodobieństwa $p(x;\theta)$ dla wszystkich stanów układu $x = (x_V, x_H) \in 0,1^H \times 0,1^V$ prawdopodobieństwo brzegowe widocznych stanów przyjmuje postać:

$$p(x_V;\theta) = \sum_{x_H \in 0,1^H} p(x;\theta),\ x_V \in \{0,1\}^V. \qquad (4)$$

- Ograniczonymi maszynami Boltzmanna nazywamy maszyny Boltzmanna, gdzie oddziaływanie pomiędzy neuronami zachodzi tylko pomiędzy neuronami z warstwy widocznej a neuronami z warstwy niewidocznej. To znaczy: $I = \{\{i, j\} : i \in i \in V, j \in H\}$.



Rysunek: `https://commons.wikimedia.org/wiki/File:Restricted_Boltzmann_machine.svg`

- Ten model sieci neuronowej jest przykładem modelu bazującego na pojęciu funkcji energetycznej. Oddziaływanie pomiędzy widocznymi x_V^i i niewidocznymi x_H^j neuronami opisuje funkcja energii:

$$\mathbf{E}(x_v^i, x_h^j) = -\sum_j b_j x_h^j - \sum_{ij} x_v^i w_{ij} x_h^j - \sum_i c_i x_v^i, \qquad (5)$$

gdzie (b_j, W_{ij}, c_i) to wagi.

- Lub w zapisie macierzowym:

$$\mathbf{E}(x_v, x_h) = y^T W x + c^T y + b^T x, \qquad (6)$$

gdzie x to stan widocznych neuronów, y niewidocznych, a Z to funkcja rozkładu.

- Zatem ostatecznie rozkład prawdopodobieństwa dla ograniczonych maszyn Boltzmanna przyjmuje postać:

$$p(x;\theta) = \frac{\sum_{y \in \{0,1\}^H} \exp(y^T W x + c^T y + b^T x)}{Z(W, c, b)}, x \in \{0,1\}^V, \tag{7}$$

gdzie, tak jak na poprzednim slajdzie, x to stan widocznych neuronów, y niewidocznych, a Z to funkcja rozkładu.

Ograniczone maszyny Boltzmanna

- Do określania rozbieżności pomiędzy rozkładem prawdopodobieństwa dla danych $P(\{v_i\})$ a rozkładem dla modelu teoretycznego używamy *dywergencji Kullbacka-Leiblera* w postaci:

$$D_{KL}(P(x_v^i) \mid\mid p_\lambda(x_v^i) = \sum_{x_v^i} P(x_v^i) \log \left(\frac{P(x_v^i)}{p_\lambda(x_v^i)} \right). \qquad (8)$$

- Zatem jeśli ograniczona maszyna Boltzmanna dokładnie odtwarza rozkład prawdopodobieństwa danych to:

$$D_{KL}(P(x_v^i) \mid\mid p_\lambda(x_v^i) = 0. \qquad (9)$$

Uczenie ograniczonych maszyn Boltzmanna

- Proces uczenia, w tym ograniczonych maszyn Boltzmanna, możemy również opisać jako

$$\min_{a,W,b} D_{KL}(P(x_v^i) \mid\mid P_\lambda(x_v^i))). \tag{10}$$

- Dywergencja KL nie jest metryką. Nie jest bowiem symetryczna:

$$D_{KL}(P \mid\mid Q)) \neq D_{KL}(Q \mid\mid P)), \tag{11}$$

ani nie spełnia nierówności trójkąta:

$$D_{KL}(R \mid\mid P)) \leqslant D_{KL}(P \mid\mid Q)) + D_{KL}(R \mid\mid Q)). \tag{12}$$

- Pytanie: czy istnieją lepsze alternatywy?

Problem transportu optymalnego

- Problem **transportu optymalnego** (ang. optimal transport) po raz pierwszy sformułował Gaspad Monge w 1781 roku.
- Następnie latach dwudziestych (dwudziestego wieku) problem został sformalizowany przez A.N. Tolstoi.
- Najczęściej w literaturze problem występuje pod hasłem **zagadnienia transportowego.**

Rysunek: Gaspard Monge, 1746-1818

Odległość Wassersteina

Definicja metryki Wassersteina

- Odległość Wassersteina, niech (X, d) będzie *przestrzenią metryczną polską* i niech $p \in [1, \infty)$. Dla każdych dwóch miar probabilistycznych, μ, i ν, na X **odległością Wassersteina** stopnia p pomiędzy tymi miarami, definiujemy następująco:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_X d(x,y)^p d\pi(x,y) \right)^{\frac{1}{p}}$$

$$= inf\{[\mathbb{E}d(X,Y)^p]^{\frac{1}{p}}.\}$$

Definicja γ-gładkiej metryki Wassersteina

- Zakładamy dwa prawdopodobieństwa p i q ze zbioru $X = \{0,1\}^d$, mamy funkcję $D : X \times X \to \mathbb{R}_+$, a α i β to dwie funkcje na X. Wtedy niech γ-gładka odległość Wassersteina będzie zdefiniowana w następujący sposób:

$$W_(p,q)_\gamma = \max_{\alpha,\beta}[\mathbb{E}_X[\alpha(X)] + \mathbb{E}_{X'}[\beta(X')]$$

$$-\gamma \sum_{x,x' \in \{0,1}^{d} \exp(\frac{1}{\gamma}(\alpha(x) + \beta(x') - D(x,x')) - 1)].$$

- Podobnie jak w przypadku dywergencji Kullbacka-Leiblera proces uczenia można sformułować jako:

$$\min_\theta W_\gamma(p_{data}, p_\theta). \tag{13}$$

- Mając prawdopodobieństwo dla ograniczonych maszyn Boltzmanna::

$$p(x;\theta) = \frac{\sum_{y\in\{0,1\}^H} \exp(y^T W x + c^T y + b^T x)}{Z(W,c,b)}, x \in \{0,1\}^V, \tag{14}$$

gdzie, tak jak na poprzednim slajdzie, x to stan widocznych neuronów, y niewidocznych, a Z to funkcja rozkładu.

- oraz stosując współczynnik regularyzacji:

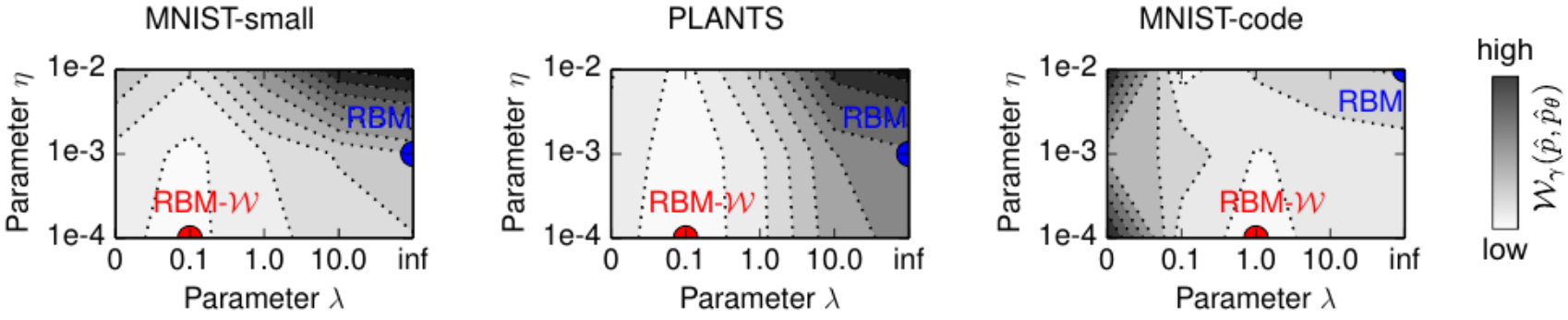
$$\Omega(c,W,b) = KL(p_{dat}||p_{c,W,b}) + \eta\left(\|a\|^2 + \sum_j \|W_j\|^2\right).$$

- Zatem ostatecznie problem uczenia można sformułować jako:

$$\min_{c,W,b} W_\gamma(p_{c,W,b}, p_{data}) + \lambda\Omega(c,W,b), \tag{15}$$

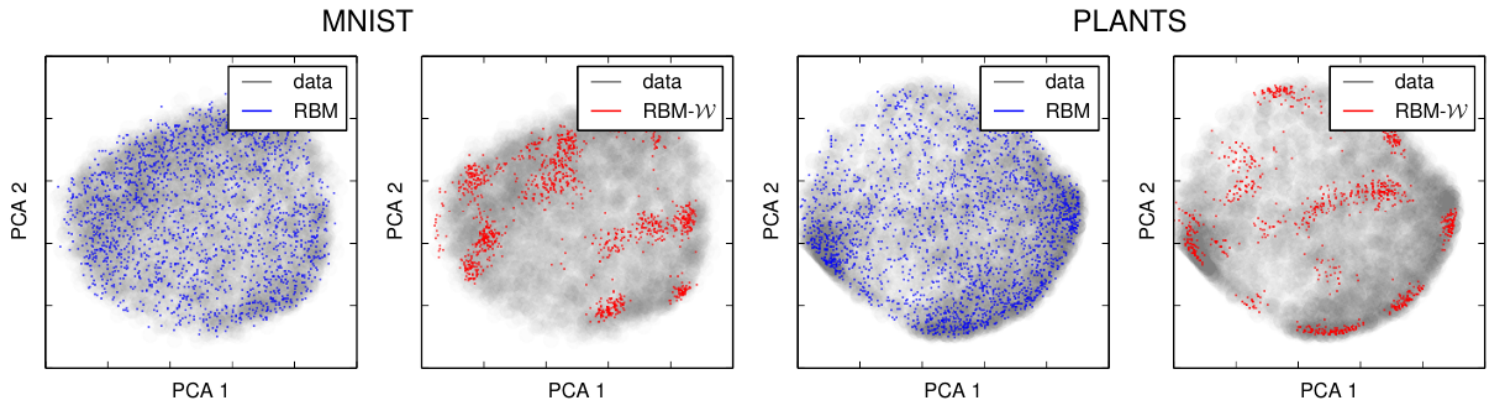
gdzie η i γ to współczynniki regularyzacji.

Wyniki empiryczne



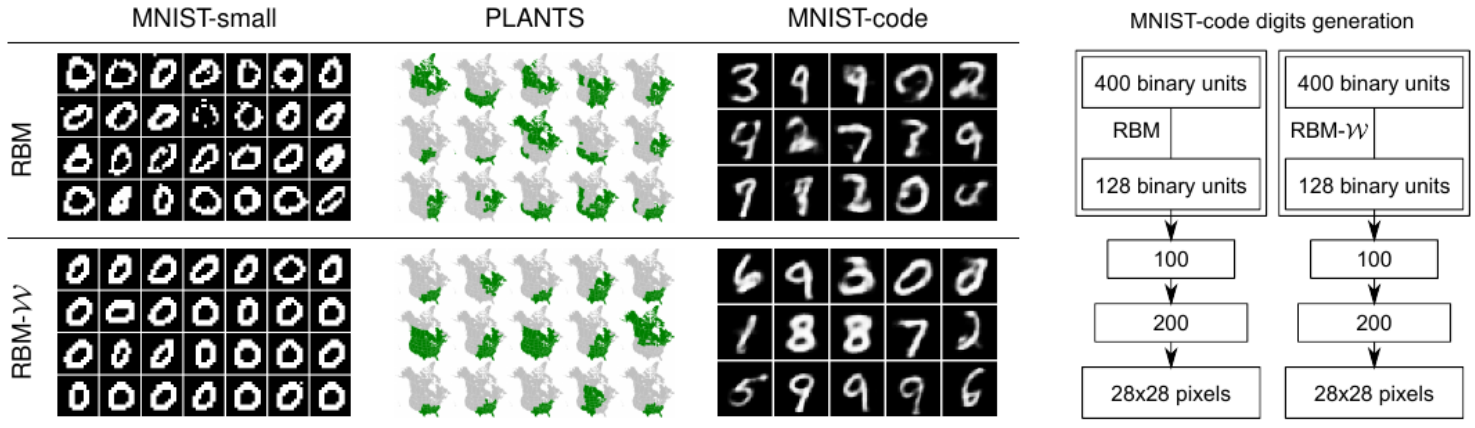
Rysunek: Montavon, Grégoire and Müller, Klaus-Robert and Cuturi, Marco. "Wasserstein Training of Restricted Boltzmann Machines", Advances in Neural Information Processing Systems 29.

Wyniki empiryczne



Rysunek: Montavon, Grégoire and Müller, Klaus-Robert and Cuturi, Marco. "Wasserstein Training of Restricted Boltzmann Machines", Advances in Neural Information Processing Systems 29.

Wyniki empiryczne



Rysunek: Montavon, Grégoire and Müller, Klaus-Robert and Cuturi, Marco. "Wasserstein Training of Restricted Boltzmann Machines", Advances in Neural Information Processing Systems 29.

Podsumowanie

Dziękuje za uwagę


Montúfar, Guido. Żestricted Boltzmann Machines: Introduction and Review."Information Geometry and its Applications IV. Springer, Cham, 2016.


Montavon, Grégoire and Müller, Klaus-Robert and Cuturi, Marco. "Wasserstein Training of Restricted Boltzmann Machines", Advances in Neural Information Processing Systems 29.


Li Wuchen, and Guido Montúfar. Ńatural gradient via optimal transport."Information Geometry 1.2 (2018): 181-214.