Иерархические модели

Московский Физико-Технический Институт

2022

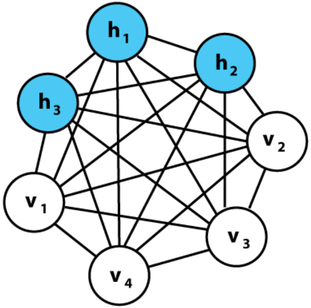
Машина Больцмана

Energy-based модель: $p(\boldsymbol{x}) = \frac{\exp(-E(\boldsymbol{x}))}{Z}$,

$$E(\boldsymbol{x}) = -\boldsymbol{x}^{\top}\boldsymbol{W}\boldsymbol{x} - \boldsymbol{w}_b^{\top}\boldsymbol{x}.$$

Вариант со скрытыми переменными:

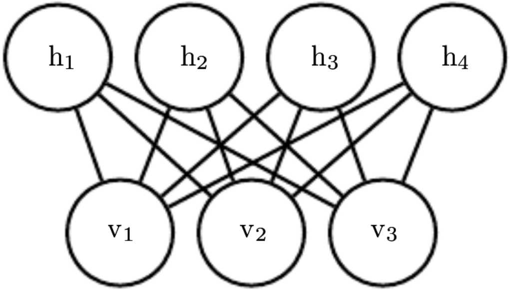
$$E(\boldsymbol{x}) = -\boldsymbol{x}^{\top}\boldsymbol{W}_v\boldsymbol{x} - \boldsymbol{x}^{\top}\boldsymbol{W}_{vh}\boldsymbol{h} - \boldsymbol{h}^{\top}\boldsymbol{W}_h\boldsymbol{h} - \boldsymbol{w}_{bh}^{\top}\boldsymbol{h} - \boldsymbol{w}_{bv}^{\top}\boldsymbol{x}.$$



Ограниченная машина Больцмана

Частный случай машины Больцмана — модель представима в виде двудольного графа:

$$E(\boldsymbol{x}) = -\boldsymbol{h}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{h} - \boldsymbol{w}_{bh}^{\mathsf{T}}\boldsymbol{h} - \boldsymbol{w}_{bb}^{\mathsf{T}}\boldsymbol{x}.$$

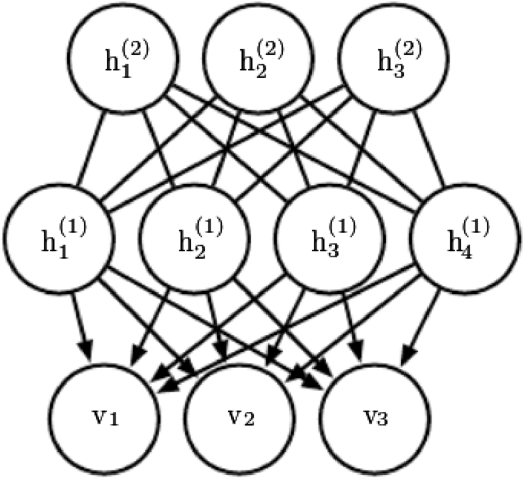


Свойства RBM

- Модель ненаправленная
- $p(\boldsymbol{h}|\boldsymbol{x}) = \prod_{j=1}^{n_h} \sigma\left((2\boldsymbol{h} - 1) \cdot (\boldsymbol{w}_{bh} + \boldsymbol{W}^\top \boldsymbol{x})\right)_j$
- $p(\boldsymbol{x}|\boldsymbol{h}) = \prod_{j=1}^{n} \sigma\left((2\boldsymbol{x} - 1) \cdot (\boldsymbol{w}_{bv} + \boldsymbol{W}\boldsymbol{h})\right)_j$

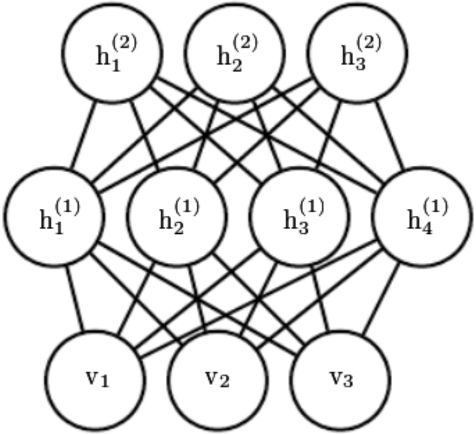
Глубокие сети доверия

- Стэк из нескольких RBM
- Оптимизируется послойно
- Модель направленная

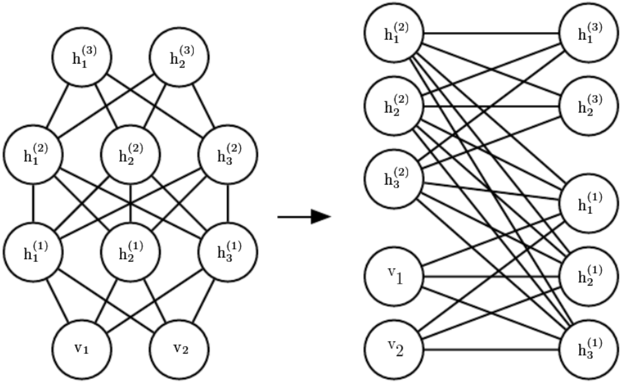


Глубокие машины Больцмана

- RBM с одним видимым слоем и несколькими скрытыми
- Модель ненаправленная

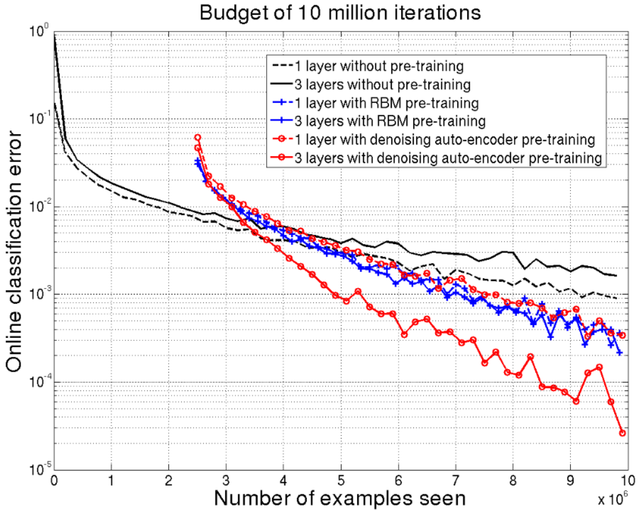


Глубокие машины Больцмана

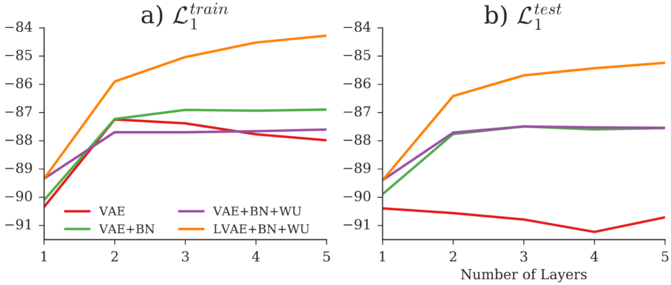


Обуславливаемся на одну долю графа $\rightarrow$ получаем независимые переменные во второй доле.

Жадное послойное обучение моделей



Жадное послойное обучение моделей: не всегда работает



Contrastive Divergence: идея

Energy-based model:

$$p(x|w) = \frac{\exp(-E_w(x))}{Z(w)}, \quad Z = \int_x \exp(-E_w(x)),$$

$$\frac{\partial \log p(x|w)}{\partial w} = \mathsf{E}_{x'} \frac{\partial E(x')}{\partial w} - \frac{\partial E(x)}{\partial w}$$

Алгоритм для RBM:

- Берем x из выборки
- $h_0 \sim p(h_0|x)$
- $x_1 \sim p(x|h_0)$
- ...
- Получаем x_k
- $\frac{\partial \log p(x|w)}{\partial w} = \frac{\partial E(x_k)}{\partial w} - \frac{\partial E(x)}{\partial w}$

Дискриминативная модель как EBM

Our key observation in this work is that one can slightly re-interpret the logits obtained from f_θ to define $p(\mathbf{x}, y)$ and $p(\mathbf{x})$ as well. Without changing f_θ, one can re-use the logits to define an energy based model of the joint distribution of data point $\mathbf{x}$ and labels y via:

$$p_\theta(\mathbf{x}, y) = \frac{\exp\left(f_\theta(\mathbf{x})[y]\right)}{Z(\theta)}, \tag{5}$$

where $Z(\theta)$ is the unknown normalizing constant and $E_\theta(\mathbf{x}, y) = -f_\theta(\mathbf{x})[y]$.

By marginalizing out y, we obtain an unnormalized density model for $\mathbf{x}$ as well,

$$p_\theta(\mathbf{x}) = \sum_y p_\theta(\mathbf{x}, y) = \frac{\sum_y \exp\left(f_\theta(\mathbf{x})[y]\right)}{Z(\theta)}. \tag{6}$$

Notice now that the $\text{LogSumExp}(\cdot)$ of the logits of *any* classifier can be re-used to define the energy function at a data point $\mathbf{x}$ as

$$E_\theta(\mathbf{x}) = -\text{LogSumExp}_y(f_\theta(\mathbf{x})[y]) = -\log \sum\nolimits_y \exp(f_\theta(\mathbf{x})[y]). \tag{7}$$

Оптимизация:

$$\log p_\theta(\mathbf{x}, y) = \log p_\theta(\mathbf{x}) + \log p_\theta(y|\mathbf{x}).$$

второе слагаемое — обычная кросс-энтропия.

Как оптимизировать первое слагаемое?

Стохастическая динамика Ланжевена

Модификация стохастического градиентного спуска:

$$T = \mathbf{x} - \lambda \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\lambda}{2})$$

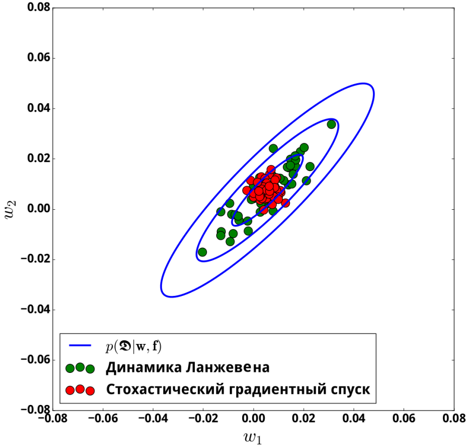
где шаг оптимизации λ изменяется с количеством итераций:

$$\sum_{\tau=1}^{\infty} \lambda_\tau = \infty, \quad \sum_{\tau=1}^{\infty} \lambda_\tau^2 < \infty.$$

Утверждение [Welling, 2011]. Распределине $T \circ T \circ \dots T$ сходится к распределению $p(\mathbf{x})$.

Стохастическая динамика Ланжевена

Распределение параметров после 2000 итераций:



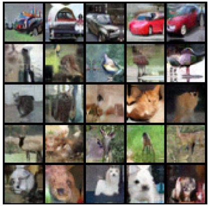
Алгоритм

Algorithm 1 JEM training: Given network f_θ, SGLD step-size α, SGLD noise σ, replay buffer B, SGLD steps η, reinitialization frequency ρ

```
while not converged do
    Sample x and y from dataset
    L_clf(θ) = xent(f_θ(x), y)
    Sample x̂_0 ~ B with probability 1 − ρ, else x̂_0 ~ U(−1, 1)          ▷ Initialize SGLD
    for t ∈ [1, 2, ..., η] do                                             ▷ SGLD
        x̂_t = x̂_{t−1} + α · ∂LogSumExp_{y'}(f_θ(x̂_{t−1})[y']) / ∂x̂_{t−1} + σ · N(0, I)
    end for
    L_gen(θ) = LogSumExp_{y'}(f(x)[y']) − LogSumExp_{y'}(f(x̂_t)[y'])    ▷ Surrogate for Eq 2
    L(θ) = L_clf(θ) + L_gen(θ)
    Obtain gradients ∂L(θ)/∂θ for training
    Add x̂_t to B
end while
```

Качество моделей

Class	Model	Accuracy% ↑	IS↑	FID↓
Hybrid	Residual Flow	70.3	3.6	46.4
	Glow	67.6	3.92	48.9
	IGEBM	49.1	8.3	**37.9**
	JEM $p(\mathbf{x}\|y)$ factored	30.1	6.36	61.8
	JEM (Ours)	**92.9**	**8.76**	38.4
Disc.	Wide-Resnet	95.8	N/A	N/A
Gen.	SNGAN	N/A	8.59	25.5
	NCSN	N/A	8.91	25.32



Литература

- Goodfellow I., Bengio Y., Courville A. Deep learning. – MIT press, 2016.
- Carreira-Perpinan M. A., Hinton G. On contrastive divergence learning //International workshop on artificial intelligence and statistics. – PMLR, 2005. – C. 33-40.
- Restricted Boltzmann Machine, a complete analysis: https://medium.com/datatype/restricted-boltzmann-machine-a-complete-analysis-part-3-contrastive-divergence-algorithm-3d06bbebb10c
- Sønderby C. K. et al. Ladder variational autoencoders //Advances in neural information processing systems. – 2016. – T. 29.
- Grathwohl W. et al. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. – 2020.
- Welling M., Teh Y. W. Bayesian learning via stochastic gradient Langevin dynamics //Proceedings of the 28th international conference on machine learning (ICML-11). – 2011. – C. 681-688.