Uczenie Maszynowe 2

Modele generatywne

dr hab. Piotr Duda, prof. AGH i PCz

Plan wykładu

Wykład 1

- **Wprowadzenia, Organiczone Maszyny Boltzmanna**

Wykład 2

- **RBM, Autoenkodery, VAE, WAE**

Wykład 3

- **GAN, Normalization Flow**

Wykład 4

- **Podejście Dyfuzyjne, Przegląd „aktualnych” modeli**

Wykład 2 – Modele generatywne

- Uczenie RBM
- Alternatywne metody uczenia
- Poza klasyczny RMB…
- …
- Autoenkodery
- Autoenkodery Wariacyjne
- Odległość Wassersteina

Organiczona Maszyna Boltzmanna

ang. Restricted Boltzmann Machine (RBM)

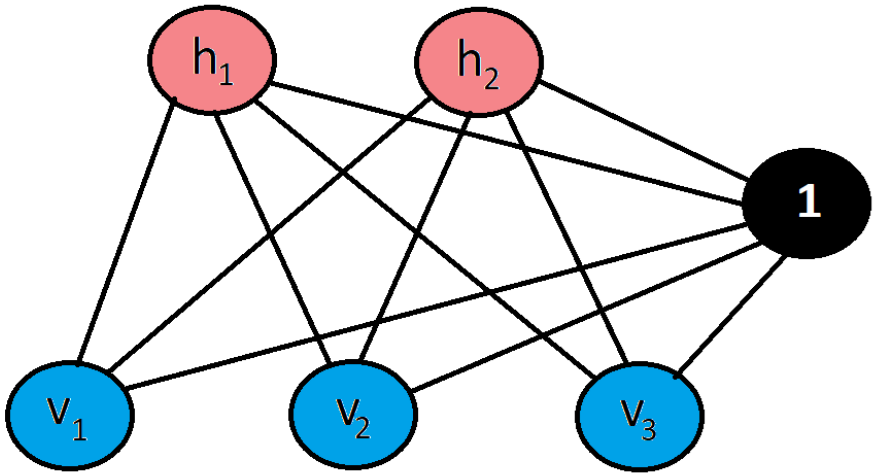
Czym są RBM?

Warstwa widoczna: $\boldsymbol{v} = \{v_1, \dots, v_d\}$

Warstwa ukryta: $\boldsymbol{h} = \{h_1, \dots, h_H\}$

Binary values: $(\boldsymbol{v}, \boldsymbol{h}) \in \{0,1\}^{d+H}$

Parametry modelu:
w – macierz wag o wymiarach d na H
b – bias warstwy widocznej, wektor długości d
c – bias warstwy ukrytej, wektor długości H



Czym są RBM?

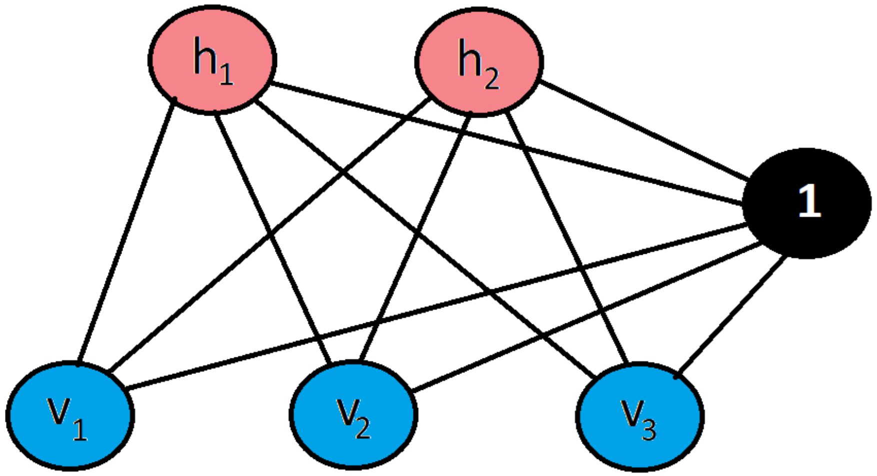
$$p(h_j = 1|\boldsymbol{v}) = \sigma\left(c_j + \sum_{i=1}^{d} w_{ij} v_i\right)$$

$$p(v_i = 1|\boldsymbol{h}) = \sigma\left(b_i + \sum_{j=1}^{H} w_{ij} h_j\right)$$

Warstwa widoczna: $\boldsymbol{v} = \{v_1, \ldots, v_d\}$

Warstwa ukryta: $\boldsymbol{h} = \{h_1, \ldots, h_H\}$

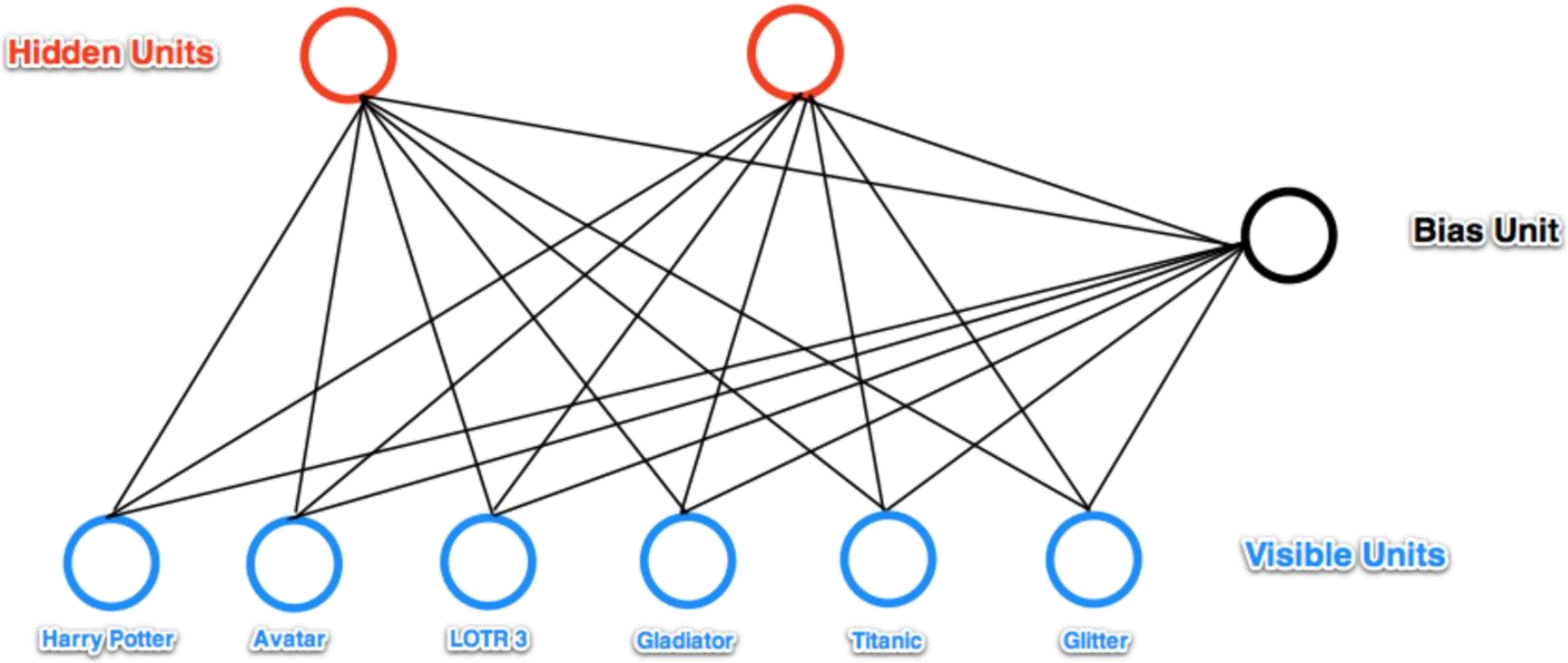
Binary values: $(\boldsymbol{v}, \boldsymbol{h}) \in \{0,1\}^{d+H}$



Model wyznacza **prawdopodobieństwa** aktywowania neuronu w danej warstwie na podstawie aktywacji neuronów w przeciwnej warstwie i parametrów modelu.

σ to funkcja sigmoidalna.

Przykład



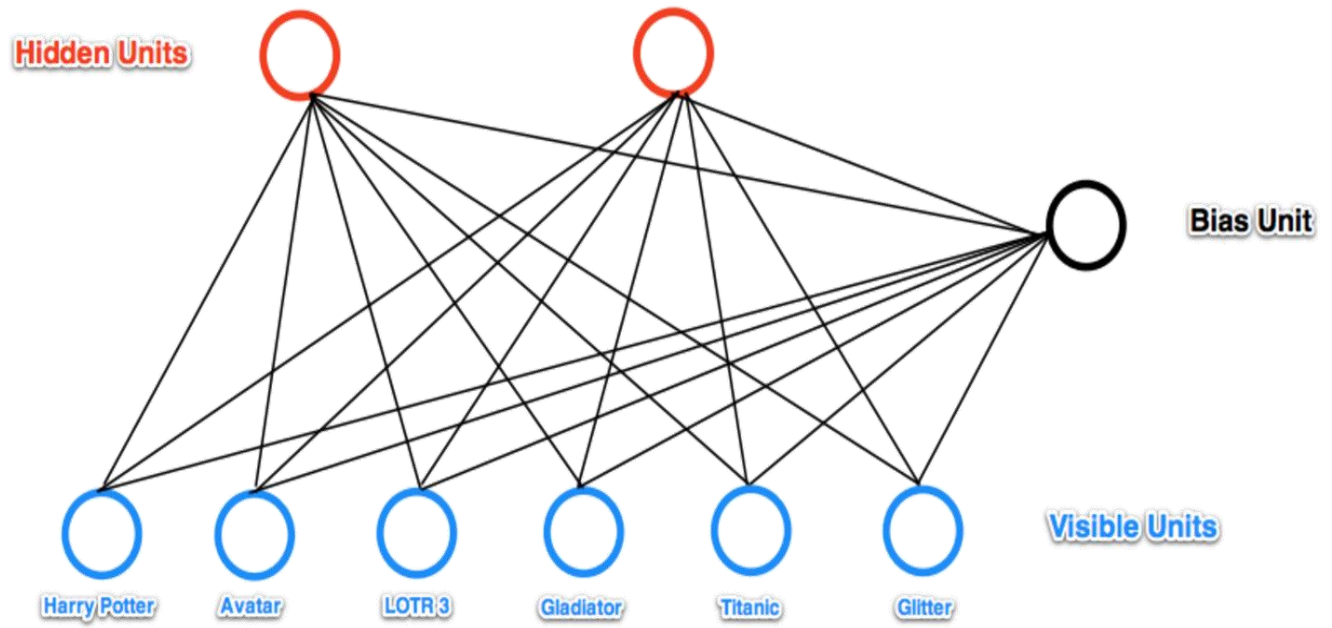
Modelowanie preferencji użytkowników

Przykład pochodzi z bloga: http://blog.echen.me/2011/07/18/introduction-to-restricted-boltzmann-machines/

Przykład

FAKE DATA:

- Alice: (Harry Potter = 1, Avatar = 1, LOTR 3 = 1, Gladiator = 0, Titanic = 0, Glitter = 0). Big SF/fantasy fan.
- Bob: (Harry Potter = 1, Avatar = 0, LOTR 3 = 1, Gladiator = 0, Titanic = 0, Glitter = 0). SF/fantasy fan, but doesn't like Avatar.
- Carol: (Harry Potter = 1, Avatar = 1, LOTR 3 = 1, Gladiator = 0, Titanic = 0, Glitter = 0). Big SF/fantasy fan.
- David: (Harry Potter = 0, Avatar = 0, LOTR 3 = 1, Gladiator = 1, Titanic = 1, Glitter = 0). Big Oscar winners fan.
- Eric: (Harry Potter = 0, Avatar = 0, LOTR 3 = 1, Gladiator = 1, Titanic = 1, Glitter = 0). Oscar winners fan, except for Titanic.
- Fred: (Harry Potter = 0, Avatar = 0, LOTR 3 = 1, Gladiator = 1, Titanic = 1, Glitter = 0). Big Oscar winners fan.



Przykład pochodzi z bloga: http://blog.echen.me/2011/07/18/introduction-to-restricted-boltzmann-machines/

Przykład

FAKE DATA:

- Alice: (Harry Potter = 1, Avatar = 1, LOTR 3 = 1, Gladiator = 0, Titanic = 0, Glitter = 0). Big SF/fantasy fan.
- Bob: (Harry Potter = 1, Avatar = 0, LOTR 3 = 1, Gladiator = 0, Titanic = 0, Glitter = 0). SF/fantasy fan, but doesn't like Avatar.
- Carol: (Harry Potter = 1, Avatar = 1, LOTR 3 = 1, Gladiator = 0, Titanic = 0, Glitter = 0). Big SF/fantasy fan.
- David: (Harry Potter = 0, Avatar = 0, LOTR 3 = 1, Gladiator = 1, Titanic = 1, Glitter = 0). Big Oscar winners fan.
- Eric: (Harry Potter = 0, Avatar = 0, LOTR 3 = 1, Gladiator = 1, Titanic = 1, Glitter = 0). Oscar winners fan, except for Titanic.
- Fred: (Harry Potter = 0, Avatar = 0, LOTR 3 = 1, Gladiator = 1, Titanic = 1, Glitter = 0). Big Oscar winners fan.

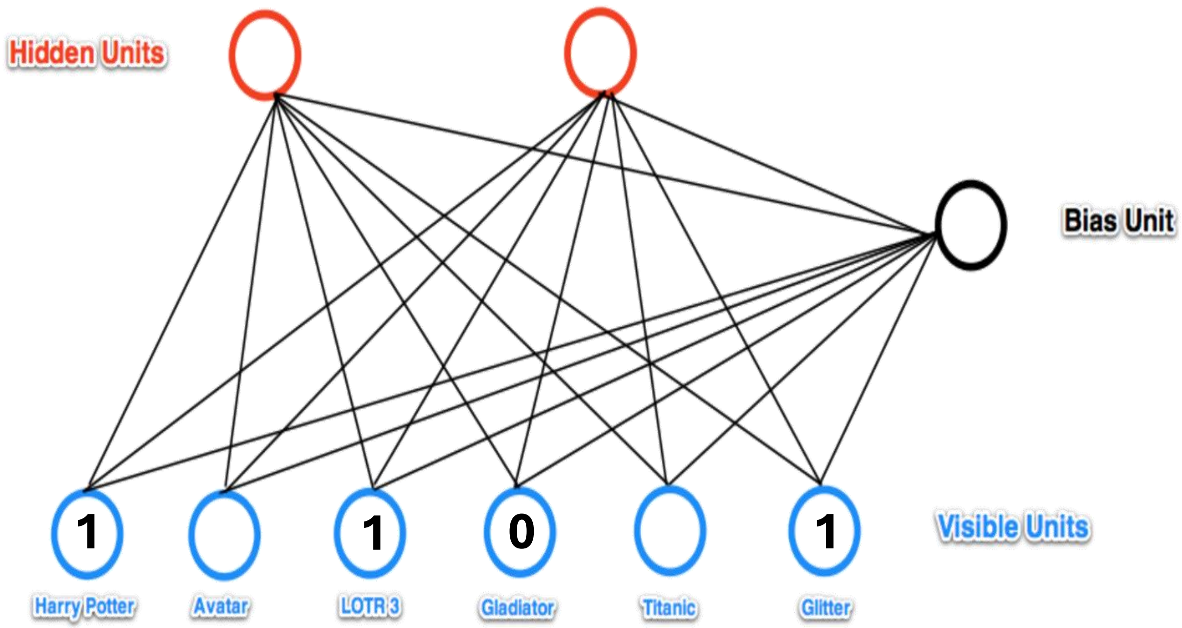
RESULT:

	Bias Unit	Hidden 1	Hidden 2
Harry Potter	-0.82602559	-7.08986885	4.96606654
Avatar	-1.84023877	-5.18354129	2.27197472
LOTR 3	3.92321075	2.51720193	4.11061383
Gladiator	0.10316995	6.74833901	-4.00505343
Titanic	-0.97646029	3.25474524	-5.59606865
Glitter	-4.44685751	-2.81563804	-2.91540988
		Big Oscar winners fan	Big SF/Fantasy fan

Przykład pochodzi z bloga: http://blog.echen.me/2011/07/18/introduction-to-restricted-boltzmann-machines/

Jak to działa?

- Naprzemiennie aktualizuj warstwy przez długi czas i generuj dane podobne do danych treningowych
- Ustaw jedną ze zmiennych ukrytych na 1 i generuj pozycje dla wybranej grupy danych
- Umieść dane z brakującymi wartościami i spróbuj je uzupełnić



Kryteria uczenia

$\boldsymbol{\theta}$ **–** parametry modelu

$\boldsymbol{S} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_N}\}$ **-** zbiór uczący

Maksymalizujemy funkcję wiarygodności dla naszego modelu względem zbioru uczącego:

$$L(\boldsymbol{\theta}|\boldsymbol{S}) = p(\boldsymbol{x_1}, \dots, \boldsymbol{x_N}|\boldsymbol{\theta})$$

Jako, że dane są niezależne to

$$L(\boldsymbol{\theta}|\boldsymbol{S}) = \prod_i p(\boldsymbol{x_i}|\boldsymbol{\theta})$$

Przekształcenia

$$\max_{\boldsymbol{\theta}}\{L(\boldsymbol{\theta}|\boldsymbol{S})\} = \max_{\boldsymbol{\theta}}\{\ln L(\boldsymbol{\theta}|\boldsymbol{S})\}$$

$$\ln L(\boldsymbol{\theta}|\boldsymbol{S}) = \sum_i \ln p(\boldsymbol{x_i}|\boldsymbol{\theta})$$

$q(\boldsymbol{x})$ **–** true distribution of data

$$\frac{1}{N}\sum_i \ln p(\boldsymbol{x_i}|\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln p(\boldsymbol{x}|\boldsymbol{\theta})$$

$$\max_{\boldsymbol{\theta}}\{\ln L(\boldsymbol{\theta}|\boldsymbol{S})\} = \max_{\boldsymbol{\theta}}\left\{\sum_i \ln p(\boldsymbol{x_i}|\boldsymbol{\theta})\right\} = \max_{\boldsymbol{\theta}}\left\{\sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln p(\boldsymbol{x}|\boldsymbol{\theta})\right\} =$$

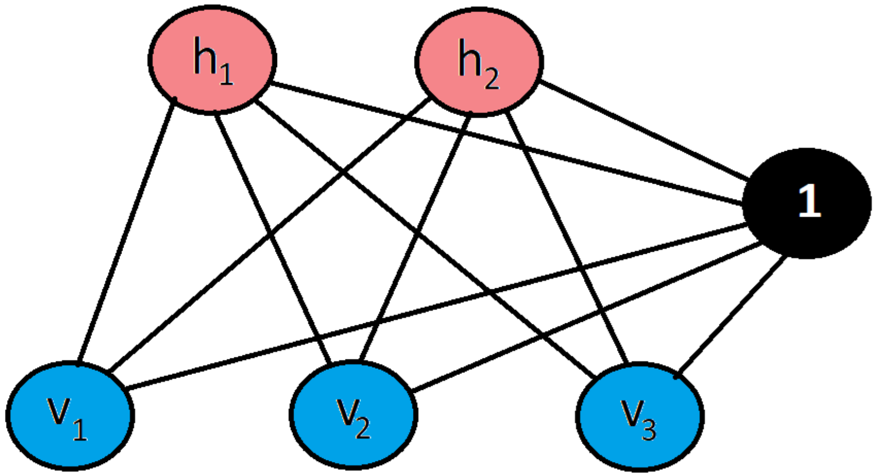
$$= \min_{\boldsymbol{\theta}}\left\{-\sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln p(\boldsymbol{x}|\boldsymbol{\theta})\right\} = \min_{\boldsymbol{\theta}}\left\{\sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln q(\boldsymbol{x}) - \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln p(\boldsymbol{x}|\boldsymbol{\theta})\right\} = \min_{\boldsymbol{\theta}}\{KL(q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{\theta})\}$$

Uczenie RMB

Warstwa widoczna: $\boldsymbol{v} = \{v_1, \dots, v_d\}$

Warstwa ukryta: $\boldsymbol{h} = \{h_1, \dots, h_H\}$

Binary values: $(\boldsymbol{v}, \boldsymbol{h}) \in \{0,1\}^{d+H}$



Łączny rozkład prawdopodpobieństwa: $p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})$

Brzegowy rozkład dla neuronów warstwy widocznej:

$$p(\boldsymbol{v}|\boldsymbol{\theta}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta})}$$

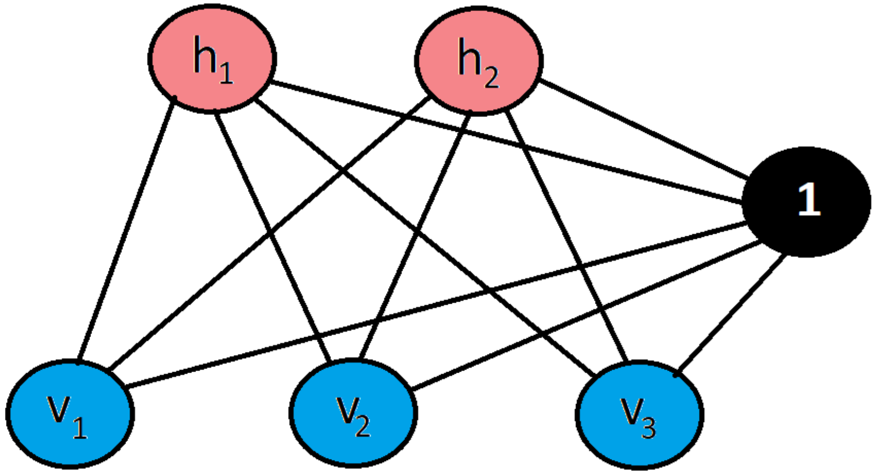
Funkcja normalizująca

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta})}$$

Gradient funkcji wiarygodności

Dla uproszczenia załóżmy, że zbiór uczący składa się tylko z jednego elementu $\hat{\boldsymbol{v}}$.
Innymi słowy $\boldsymbol{S} = \{\hat{\boldsymbol{v}}\}$

$$L(\boldsymbol{\theta}|\hat{\boldsymbol{v}}) = p(\hat{\boldsymbol{v}}|\boldsymbol{\theta}) = \frac{\tilde{p}(\hat{\boldsymbol{v}}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} = \frac{\sum_{\boldsymbol{h}} e^{-E(\hat{\boldsymbol{v}},\boldsymbol{h}|\theta)}}{\sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta})}}$$



$$\frac{\partial \ln L(\boldsymbol{\theta}|\hat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}\left(\ln \sum_{\boldsymbol{h}} e^{-E(\hat{\boldsymbol{v}},\boldsymbol{h}|\boldsymbol{\theta})}\right) - \frac{\partial}{\partial \boldsymbol{\theta}}\left(\ln \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta})}\right) =$$

$$= -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\hat{\boldsymbol{v}}) \frac{\partial E(\hat{\boldsymbol{v}},\boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{v},\boldsymbol{h}) \frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Gradient funkcji wiarygodności

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\sum\nolimits_{\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{h}|\widehat{\boldsymbol{v}}) \frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum\nolimits_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

Gradient funkcji wiarygodności

$$p(h_j = 1|\boldsymbol{v}) = \sigma\left(c_j + \sum_{i=1}^{d} w_{ij} v_i\right)$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\sum\nolimits_{\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{h}|\widehat{\boldsymbol{v}}) \frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum\nolimits_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{d} \sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{d} b_i v_i - \sum_{j=1}^{H} c_j h_j$$

$$p(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{H} p(h_j|\boldsymbol{v}) \qquad p(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{d} p(v_i|\boldsymbol{h})$$

Gradient funkcji wiarygodności

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\sum\nolimits_{\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{h}|\widehat{\boldsymbol{v}}) \frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum\nolimits_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{d}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{d} b_i\, v_i - \sum_{j=1}^{H} c_j h_j$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial w_{ij}} = -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\hat{v}_i h_j\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[v_i h_j\right]$$

Gradient funkcji wiarygodności

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\sum\nolimits_{\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{h}|\widehat{\boldsymbol{v}}) \frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum\nolimits_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{d}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{d} b_i\, v_i - \sum_{j=1}^{H} c_j h_j$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial b_i} = -\hat{v}_i + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}[v_i]$$

Gradient funkcji wiarygodności

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\sum\nolimits_{\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{h}|\widehat{\boldsymbol{v}}) \frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum\nolimits_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{d}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{d} b_i\, v_i - \sum_{j=1}^{H} c_j h_j$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial c_j} = -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}[h_j] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[h_j\right]$$

Gradient funkcji wiarygodności

$$\frac{\partial \ln L(\boldsymbol{\theta}|\hat{\boldsymbol{v}})}{\partial w_{ij}} = -\mathbb{E}_{p(\boldsymbol{h}|\hat{\boldsymbol{v}})}\left[\hat{v}_i h_j\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[v_i h_j\right]$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\hat{\boldsymbol{v}})}{\partial b_i} = -\hat{v}_i + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}[v_i]$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\hat{\boldsymbol{v}})}{\partial c_j} = -\mathbb{E}_{p(\boldsymbol{h}|\hat{\boldsymbol{v}})}\left[h_j\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[h_j\right]$$

Gradient funkcji wiarygodności

Wersja mini-batchowa: $\boldsymbol{S} = \{\boldsymbol{v_1}, \dots, \boldsymbol{v_M}\}$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\hat{\boldsymbol{v}})}{\partial w_{ij}} = -\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v_m})}\left[\boldsymbol{v_{m,i}} h_j\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[v_i h_j\right]$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\hat{\boldsymbol{v}})}{\partial b_i} = -\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{v_m} + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}[v_i]$$

$$\frac{\partial \ln L(\boldsymbol{\theta}|\hat{\boldsymbol{v}})}{\partial c_j} = -\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v_m})}\left[h_j\right] + \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[h_j\right]$$

Przybliżenie rozkładu łącznego

Próbkowanie Gibbsa:

$$p\left(h_j = 1 \middle| \boldsymbol{v}\right) = \sigma\left(c_j + \sum_{i=1}^{d} w_{ij} v_i\right)$$

$$p(v_i = 1|\boldsymbol{h}) = \sigma\left(b_i + \sum_{j=1}^{H} w_{ij} h_j\right)$$

Procedura:

1) Weź losowe $\boldsymbol{v}$
2) Uaktualnij $\boldsymbol{h}$: $p(\boldsymbol{h}|\boldsymbol{v})$
3) Uaktualnij $\boldsymbol{v}$: $p(\boldsymbol{v}|\boldsymbol{h})$
4) Uaktualnij $\boldsymbol{h}$: $p(\boldsymbol{h}|\boldsymbol{v})$
5) Uaktualnij $\boldsymbol{v}$: $p(\boldsymbol{v}|\boldsymbol{h})$

....

W pliku nie można odnaleźć części obrazu z identyfikatorem relacji rId3.

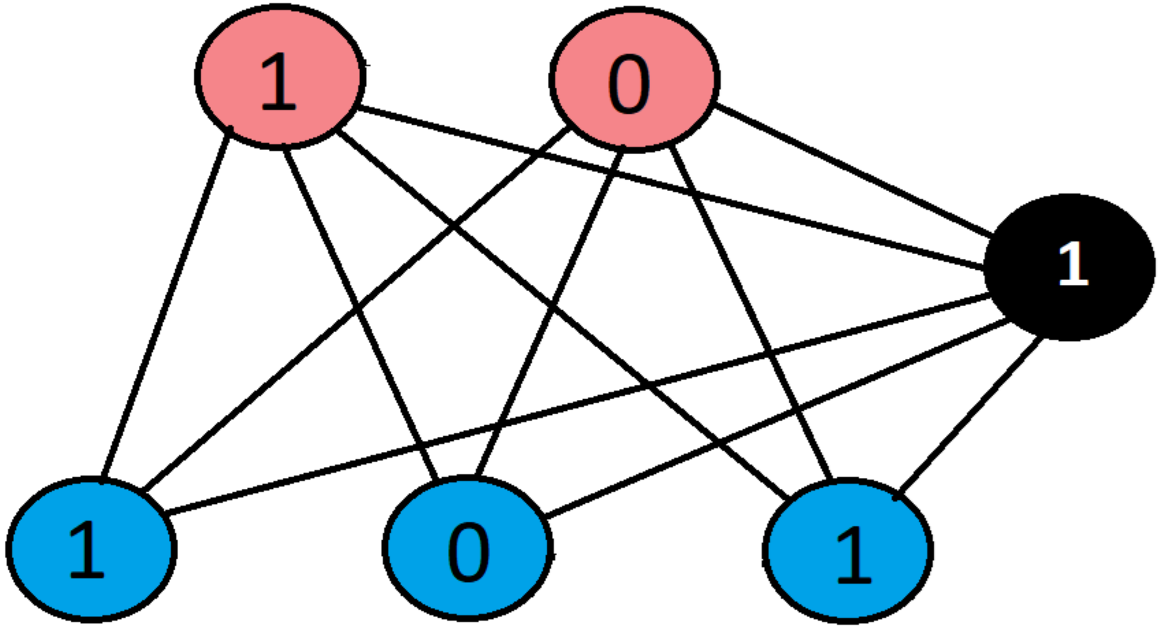
Przybliżenie rozkładu łącznego

Powtórz procedurę $\boldsymbol{Q}$ razy, gdzie k to liczba naprzemiennie wygenerowanych stanów widzialnych:

$$\left(\boldsymbol{v}_{\boldsymbol{1}}^{\boldsymbol{G},(k)}, \boldsymbol{h}_{\boldsymbol{1}}^{\boldsymbol{G},(k)}\right), \dots \left(\boldsymbol{v}_{\boldsymbol{M}}^{\boldsymbol{G},(k)}, \boldsymbol{h}_{\boldsymbol{Q}}^{\boldsymbol{G},(k)}\right)$$

Wówczas

$$\mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[v_i h_j\right] \approx \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{E}_{p\left(\boldsymbol{h}|\boldsymbol{v}_{\boldsymbol{q}}^{\boldsymbol{G},(k)}\right)}\left[v_{q,i}^{G,(k)} h_j\right]$$



Przybliżenie z użyciem próbkowania Gibbsa

Obliczony gradient

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial \boldsymbol{\theta}} = -\sum\nolimits_{\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{h}|\widehat{\boldsymbol{v}}) \frac{\partial E(\widehat{\boldsymbol{v}}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum\nolimits_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{p}(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Przybliżenie próbkowaniem Gibbsa

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial w_{ij}} \approx -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\hat{v}_i h_j\right] + \frac{1}{Q}\sum\nolimits_{q=1}^{Q} \mathbb{E}_{p\left(\boldsymbol{h}|\boldsymbol{v}_{\boldsymbol{q}}^{G,(k)}\right)}\left[v_{q,i}^{G,(k)} h_j\right]$$

Wersja mini-batchowa:

$$\frac{\partial \ln L(\boldsymbol{\theta}|\boldsymbol{S})}{\partial w_{ij}} \approx -\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v}_{\boldsymbol{m}})}\left[\boldsymbol{v}_{\boldsymbol{m},i} h_j\right] + \frac{1}{Q}\sum\nolimits_{q=1}^{Q} \mathbb{E}_{p\left(\boldsymbol{h}|\boldsymbol{v}_{\boldsymbol{q}}^{G,(k)}\right)}\left[v_{q,i}^{G,(k)} h_j\right]$$

Alternatywne metody uczenia

Contrastive divergence

$\boldsymbol{v}^{(0)}$ - wektor warstwy widocznej (ze zbioru uczącego)
$\boldsymbol{v}^{(k)}$ - wektor warstwy widocznej po k krokach Gibbs samplingu, rozpoczętego w punkcie $\boldsymbol{v}^{(0)}$

1) Weź element minibatcha $\boldsymbol{v}$
2) Oblicz aktywację warstwy ukrytej $\boldsymbol{h}$ względem $p(\boldsymbol{h}|\boldsymbol{v})$

$$p(h_j = 1|\boldsymbol{v}) = \sigma\left(c_j + \sum_{i=1}^{d} w_{ij} v_i\right)$$

3) Oblicz $v_i h_j$ jako $\mathrm{E}[v_i h_j]_{data}$
4) Powtórz k razy:
    2a) Uaktualnij $\boldsymbol{h}$: $p(\boldsymbol{h}|\boldsymbol{v})$
    2b) Uaktualnij $\boldsymbol{v}$: $p(\boldsymbol{v}|\boldsymbol{h})$
5) Oblicz $v_i h_j$ jako $\mathrm{E}[v_i h_j]_{recon}$
6) Uśrednij $\mathrm{E}[v_i h_j]_{data}$ i $\mathrm{E}[v_i h_j]_{recon}$ dla wszystkich elementów z minibatcha

Contrastive divergence

Uczenie na podstawie minibatacha ***S***

Gradient:

$$\frac{\partial \ln L(\boldsymbol{\theta}|\boldsymbol{S})}{\partial w_{ij}} = -\mathbb{E}_{data} v_i h_j + \mathbb{E}_{model} v_i h_j$$

Contrastive Divergence (CD-k):

$$\frac{\partial \ln L(\theta|\widehat{\boldsymbol{v}})}{\partial w_{ij}} = -\mathbb{E}_{p\left(\boldsymbol{h}\middle|\widehat{\boldsymbol{v}}\right)}[v_i h_j] + \mathbb{E}_{p\left(\boldsymbol{h}\middle|\widehat{\boldsymbol{v}}^{(k)}\right)}[v_i h_j]$$

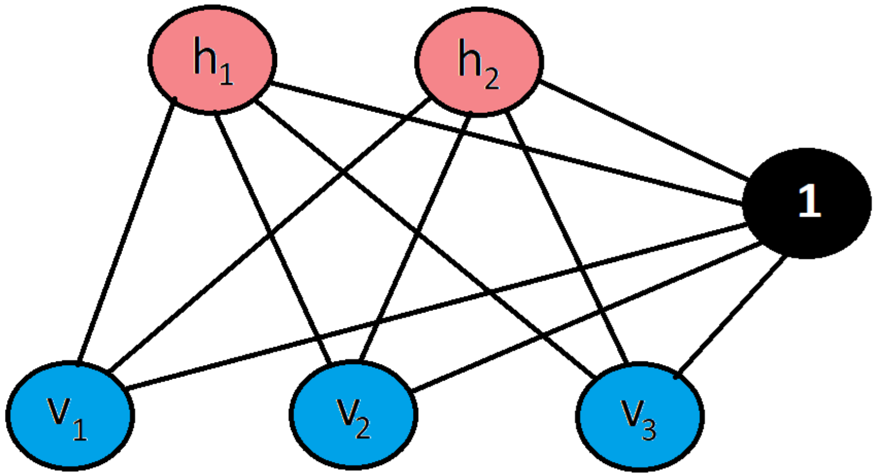
Próbkowanie Gibbsa:

$$\frac{\partial \ln L(\boldsymbol{\theta}|\widehat{\boldsymbol{v}})}{\partial w_{ij}} \approx -\mathbb{E}_{p(\boldsymbol{h}|\widehat{\boldsymbol{v}})}\left[\hat{v}_i h_j\right] + \frac{1}{Q}\sum_{q=1}^{Q} \mathbb{E}_{p\left(\boldsymbol{h}|\boldsymbol{v}_{\boldsymbol{q}}^{\boldsymbol{G},(k)}\right)}\left[v_{q,i}^{G,(k)} h_j\right]$$

Atualizacja parametrów CD:

Dla połączeń neuronów:

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \eta\left(\mathrm{E}[v_i h_j]_{recon} - \mathrm{E}[v_i h_j]_{data}\right)$$



Dla biasów:

$$b_i^{(t+1)} = b_i^{(t)} - \eta(\mathrm{E}[v_i]_{recon} - \mathrm{E}[v_i]_{data})$$

$$c_j^{(t+1)} = c_j^{(t)} - \eta\left(\mathrm{E}[h_j]_{recon} - \mathrm{E}[h_j]_{data}\right)$$

As any neural network can be learned using other methods: Momentum, AdaGrad, RMSProp, Adam

Persistent contrastive divergence

W PCD zamiast rozpoczynać od danych treningowych w każdej iteracji, utrzymujemy stały łańcuch Markowa, który kontynuuje swoją ewolucję między krokami aktualizacji parametrów modelu.
Każdy nowy gradient jest obliczany na podstawie próbek generowanych przez "persistent particles" — próbki, które pamiętają stan z poprzedniego kroku treningowego.

Inicjalizujemy *R* niezależnych łańcuchów Markowa:

$$\widetilde{\boldsymbol{v}}_{\mathbf{1}}, \dots, \widetilde{\boldsymbol{v}}_{\boldsymbol{R}}$$

W każdym kroku wykonujemy *k-krotny* Gibbs sampling dla każgo z łańcuchów:

$$\widetilde{\boldsymbol{v}}_{\mathbf{1}}^{(k)}, \dots, \widetilde{\boldsymbol{v}}_{\boldsymbol{R}}^{(k)}$$

Persistent Contrastive Divergence:

$$\frac{\partial \ln L(\theta|\boldsymbol{S})}{\partial w_{ij}} = -\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v}_{\boldsymbol{m}})}[v_i h_j] + \frac{1}{R}\sum_{r=1}^{R} \mathbb{E}_{p\left(\boldsymbol{h}|\widetilde{\boldsymbol{v}}_{\boldsymbol{r}}^{(k)}\right)}[v_i h_j]$$

Aktualizujemy łańcuchy $\widetilde{\boldsymbol{v}}_{\mathbf{1}} = \widetilde{\boldsymbol{v}}_{\mathbf{1}}^{(k)}, \dots, \widetilde{\boldsymbol{v}}_{\boldsymbol{R}} = \widetilde{\boldsymbol{v}}_{\boldsymbol{R}}^{(k)}$

Parallel tempering

Preparation:

Ustal zbiór *R* temperatur:

$$1 = T_1 < T_2 < \cdots < T_R$$

Dla każdej z temperatur utwórz osobny rozkład:

$$p_r(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z_r} e^{-\frac{1}{T_r}E(\boldsymbol{v},\boldsymbol{h})}, r = 1, \dots, R,$$

gdzie $p_1(\boldsymbol{v}, \boldsymbol{h}) = p(\boldsymbol{v}, \boldsymbol{h})$ i $Z_r = \sum_{\boldsymbol{h},\boldsymbol{v}} e^{-\frac{1}{T_r}E(\boldsymbol{v},\boldsymbol{h})}$

Inicjalizuj *R* łańcuchów Markowa wektorami z warstwy widocznej:

$$\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{1}}, \dots, \bar{\bar{\boldsymbol{v}}}_{\boldsymbol{R}}$$

Parallel tempering

Dla każdego z łańcuchów Markowa przeprowadź k krotne próbkowanie Gibbsa:

$$\left(\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{1}}^{(k)}, \bar{\bar{\boldsymbol{h}}}_{\boldsymbol{1}}^{(k)}\right), \dots, \left(\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{R}}^{(k)}, \bar{\bar{\boldsymbol{h}}}_{\boldsymbol{R}}^{(k)}\right)$$

Podmień „particles” $\left(\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{r-1}}^{(k)}, \bar{\bar{\boldsymbol{h}}}_{\boldsymbol{r-1}}^{(k)}\right)$ i $\left(\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{r}}^{(k)}, \bar{\bar{\boldsymbol{h}}}_{\boldsymbol{r}}^{(k)}\right)$ dla dwóch kolejnych łańcuchów Markowa z prawdopodobieństwem:

$$\min\left\{1, \exp\left(\left(\frac{1}{T_r} - \frac{1}{T_{r-1}}\right)\left(E\left(\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{r}}^{(k)}, \bar{\bar{\boldsymbol{h}}}_{\boldsymbol{r}}^{(k)}\right) - E\left(\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{r-1}}^{(k)}, \bar{\bar{\boldsymbol{h}}}_{\boldsymbol{r-1}}^{(k)}\right)\right)\right)\right\}$$

Po podmianie weź warstwę widoczną z pierwszego łańcucha:

$$\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{1}}^{(k)}$$

Parallel tempering

Ustaw $\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{r}} = \bar{\bar{\boldsymbol{v}}}_{\boldsymbol{r}}^{(k)}$ dla wszystkich $r = 1, \dots, R$**,** i powtórz procedurę $\boldsymbol{L}$ razy

$$\bar{\bar{\boldsymbol{v}}}_{\mathbf{1,1}}^{(k)}, \dots, \bar{\bar{\boldsymbol{v}}}_{\mathbf{1,}\boldsymbol{L}}^{(k)}$$

Parallel tempering:

$$\frac{\partial \ln L(\theta|\boldsymbol{S})}{\partial w_{ij}} = -\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p\left(\boldsymbol{h}\middle|\boldsymbol{v}_{\boldsymbol{m}}\right)}[v_i h_j] + \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{p\left(\boldsymbol{h}\middle|\bar{\bar{\boldsymbol{v}}}_{\mathbf{1,}\boldsymbol{l}}^{(k)}\right)}[v_i h_j]$$

Porównanie metod uczenia

Metoda	Zalety	Wady
Gibbs Sampling	- Prostota implementacji. - Teoretycznie dokładny w nieskończonym czasie (znajduje prawdziwy rozkład równowagi).	- Kosztowny obliczeniowo przy dużej liczbie zmiennych. - Wolna zbieżność do równowagi. - W praktyce często nieosiągalny w ograniczonym czasie.
Contrastive Divergence (CD)	- Prostota implementacji. - Szybki dzięki krótkim łańcuchom Markowa (np. CD-1).	- Niedokładne przybliżenie rozkładu równowagi. - Nieefektywny w modelach o złożonych rozkładach.
Persistent Contrastive Divergence (PCD)	- Lepsze przybliżenie rozkładu równowagi dzięki stałym łańcuchom Markowa.	- Wolniejszy niż CD (utrzymywanie łańcucha wymaga więcej obliczeń).
Parallel Tempering (PT)	- Efektywnie unika lokalnych minimów dzięki równoległym łańcuchom w różnych temperaturach. - - Umożliwia dokładne próbkowanie z rozkładu równowagi nawet w złożonych modelach.	- Bardzo kosztowny obliczeniowo (potrzebuje wielu równoległych łańcuchów). - - Wymaga dostrojenia parametrów, takich jak liczba łańcuchów i zakres temperatur.

Poza klasyczny RBM…

- ***Deep belief network***

Hinton, Geoffrey E. "Deep belief networks." Scholarpedia 4.5 (2009): 5947.

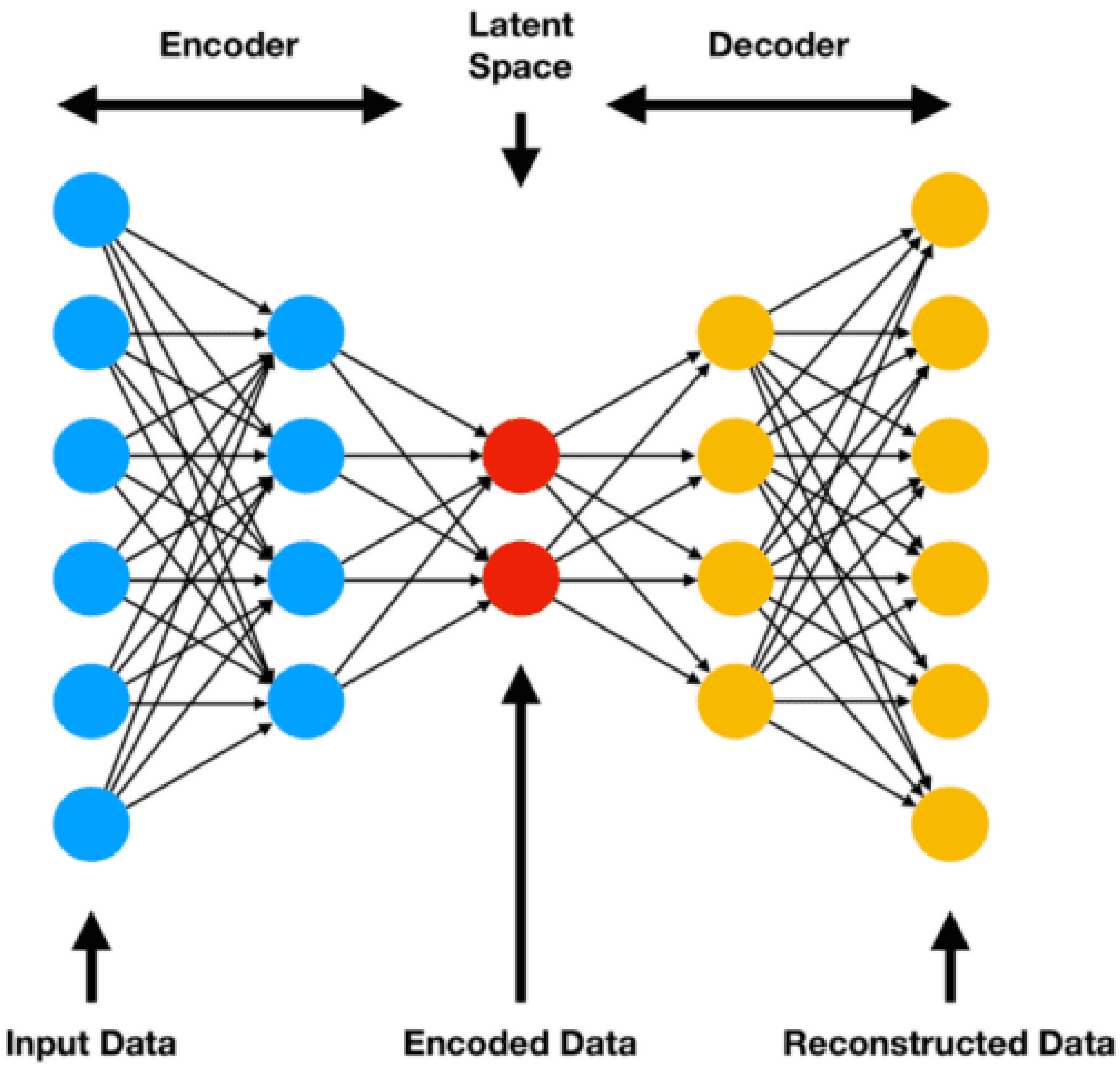
- ***Gaussian-Bernoulli RBM***

Liao, Renjie, et al. "Gaussian-bernoulli rbms without tears." arXiv preprint arXiv:2210.10318 (2022).

- ***Urestricted Boltzmann Machines***

…

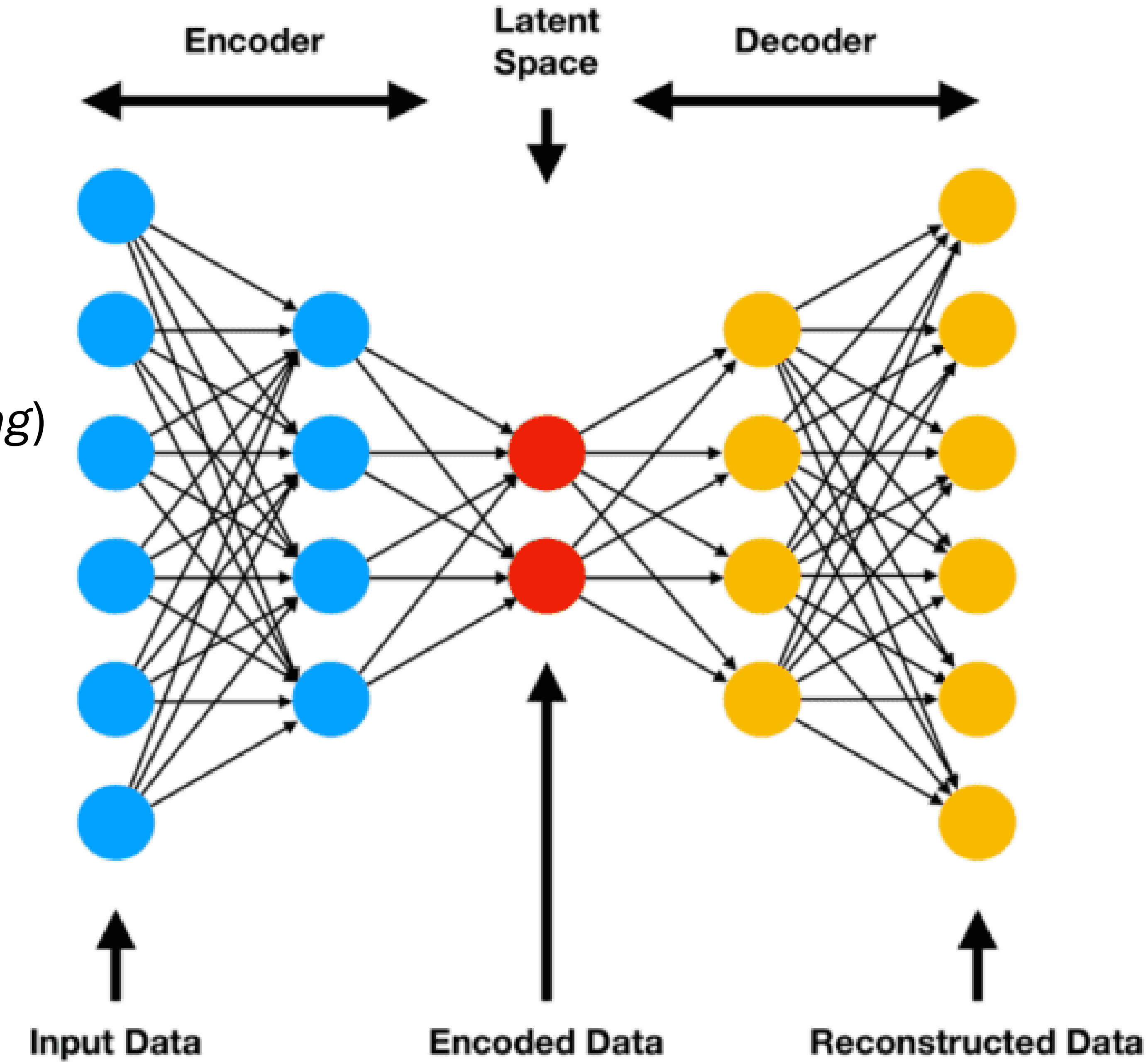
Autoekodery



Źródło: https://medium.com/ml-cheat-sheet/comprehensive-introduction-to-autoencoders-665502a75e1a

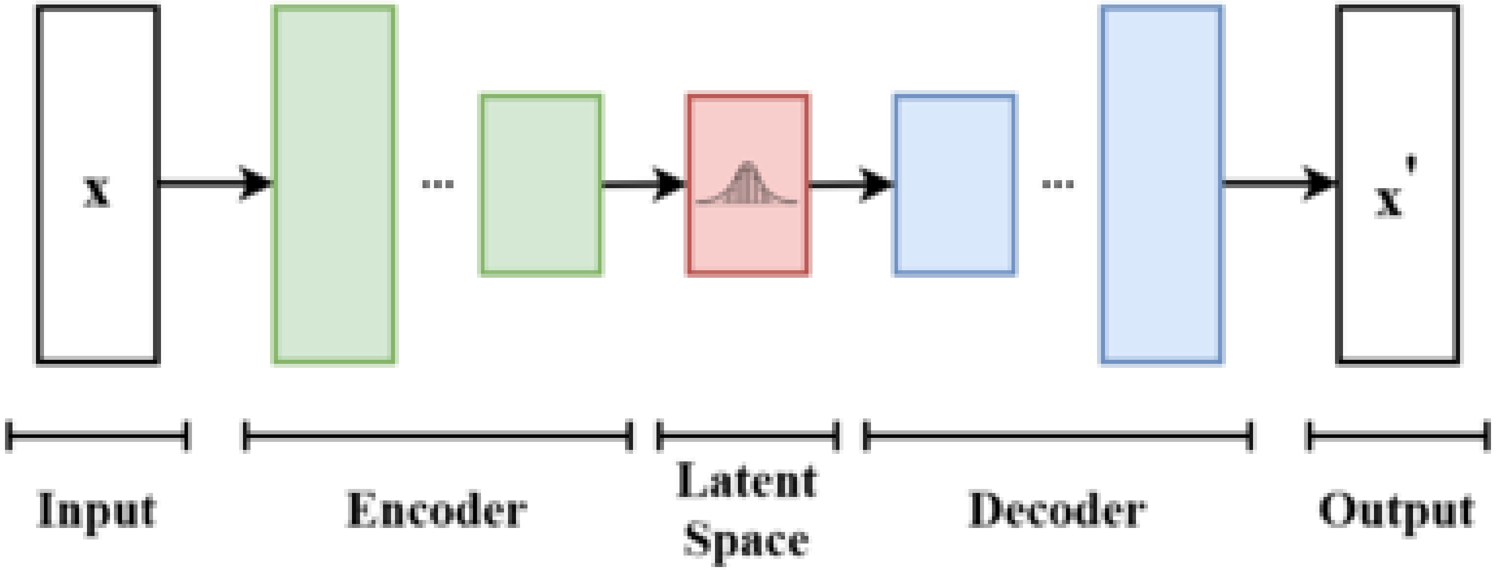
Autoekodery

- Odszumiające (*ang. Denoising*)
- Rzadkie (*ang. Sparse*)
- Contractive
- Wariacyjne (*ang. Variational*)

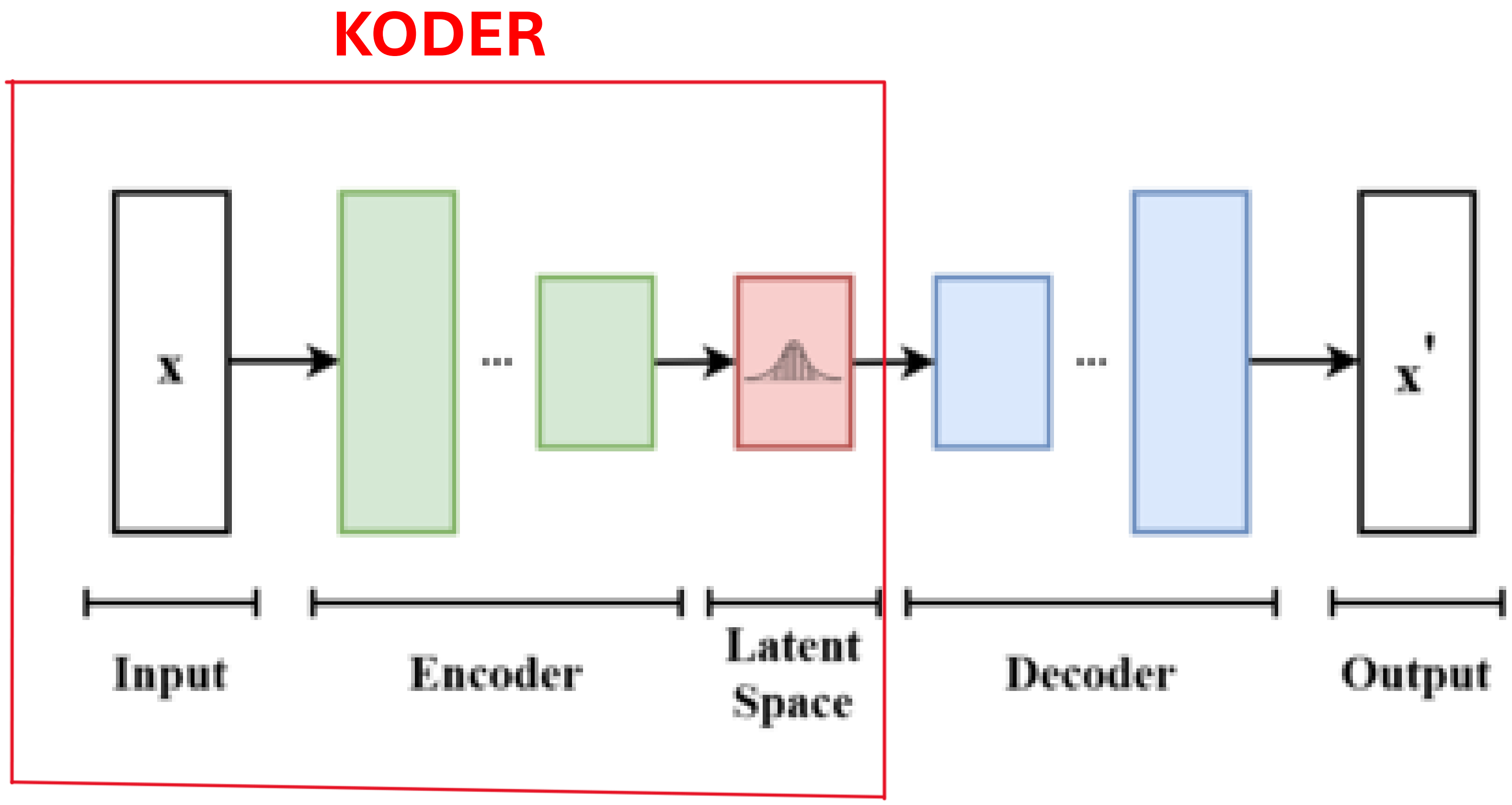


Źródło: https://medium.com/ml-cheat-sheet/comprehensive-introduction-to-autoencoders-665502a75e1a

Autoenkoder Wariacyjny



Autoenkoder Wariacyjny

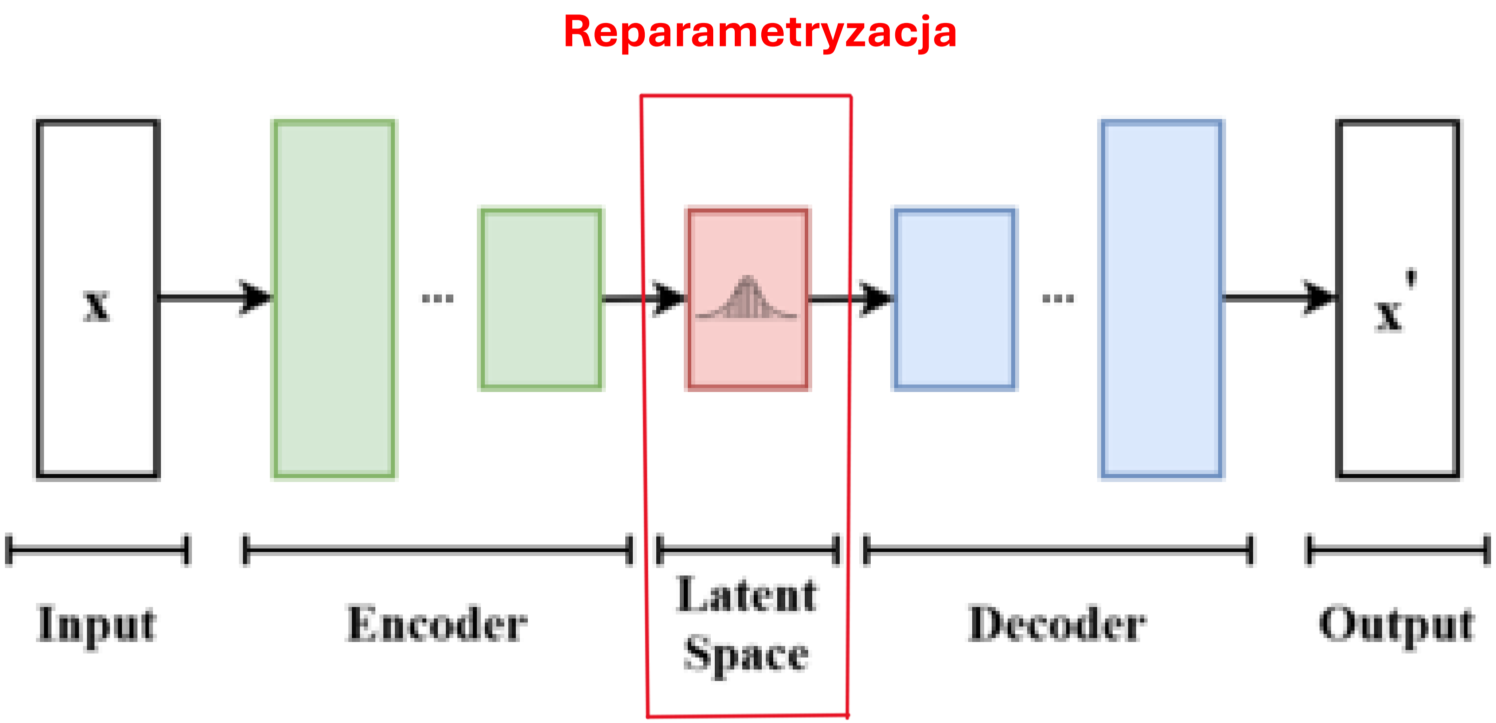


Przestrzeń latentna (*ang. Latent Space*)

Przestrzeń latentna to abstrakcyjna, niskowymiarowa przestrzeń, w której chcemy reprezentować dane wejściowe. Dzięki tej probabilistycznej reprezentacji możliwe jest:

- **Interpolacja**: Płynne przejścia między różnymi próbkami.
- **Ekstrapolacja**: Generowanie nowych danych poprzez eksplorację tej przestrzeni.
- **Zrozumienie struktury danych**: Odkrywanie ukrytych zależności i cech w danych.

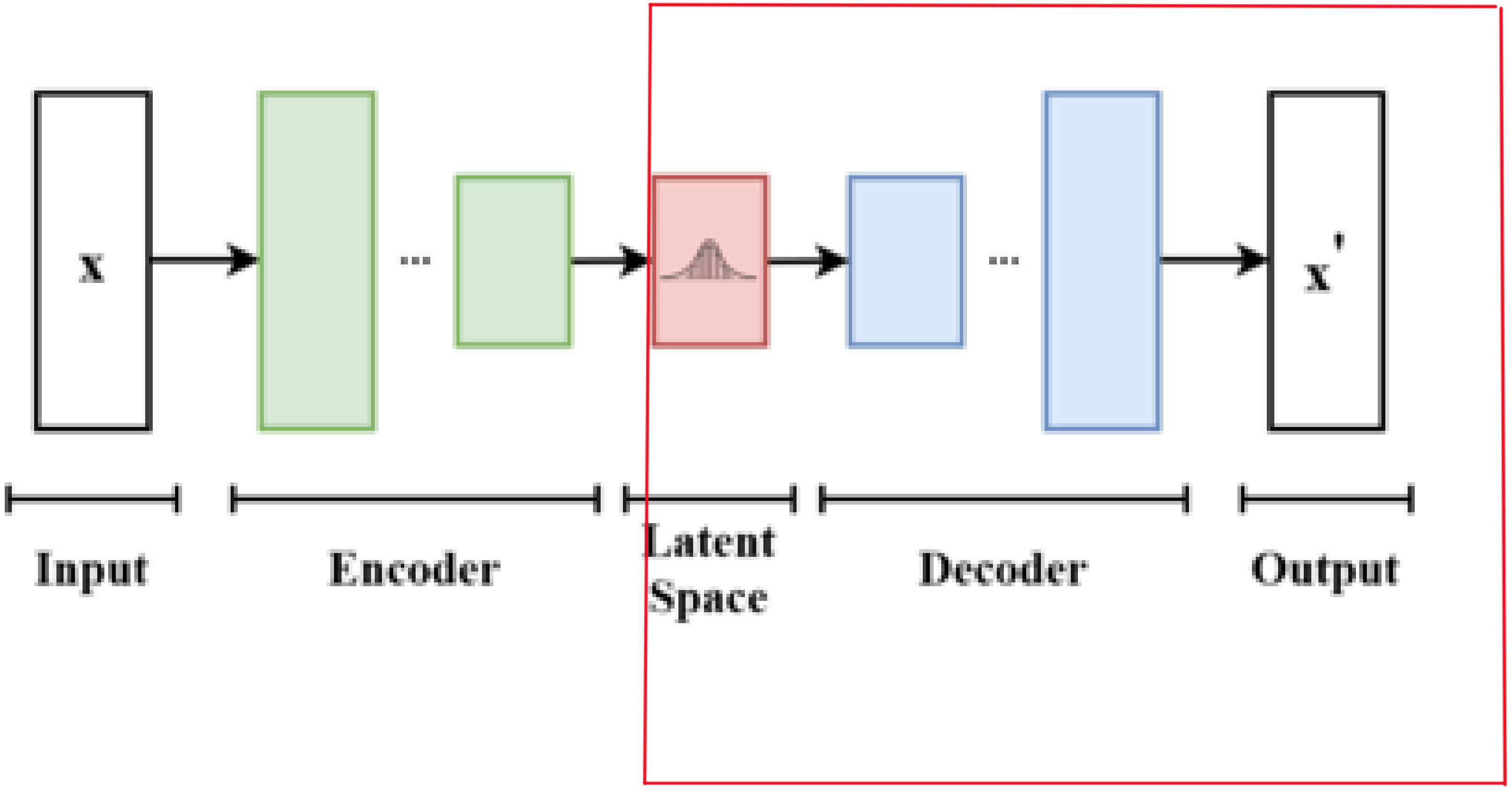
Autoenkoder Wariacyjny



$$z = \mu + \sigma \cdot \varepsilon$$

Autoenkoder Wariacyjny

DEKODER



Funkcja kosztu

Funkcja straty w VAE składa się z dwóch składników:

1. Strata rekonstrukcji

$$L_R = -\mathbb{E}_{q(z|x)} \log p(x|z)$$

$$p(x|z) = -\frac{1}{2}\|x - DEC(z)\|_2^2$$

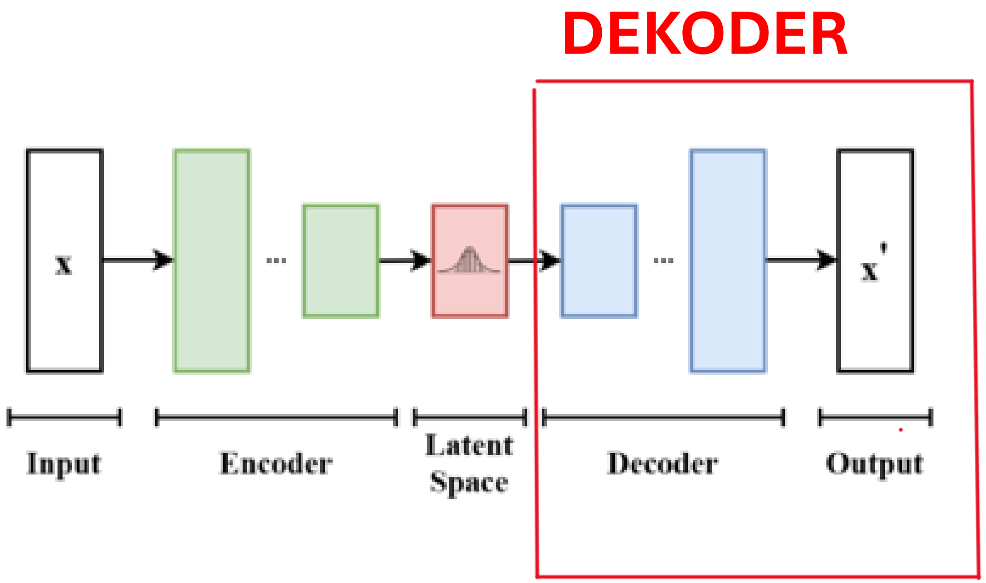
2. Dywergencja Kullbacka-Leiblera

$$L_{KL} = -KL(q(z|x)|p(z)) = -\int q\,(z|x) log \frac{q(z|x)}{p(z)} dz$$

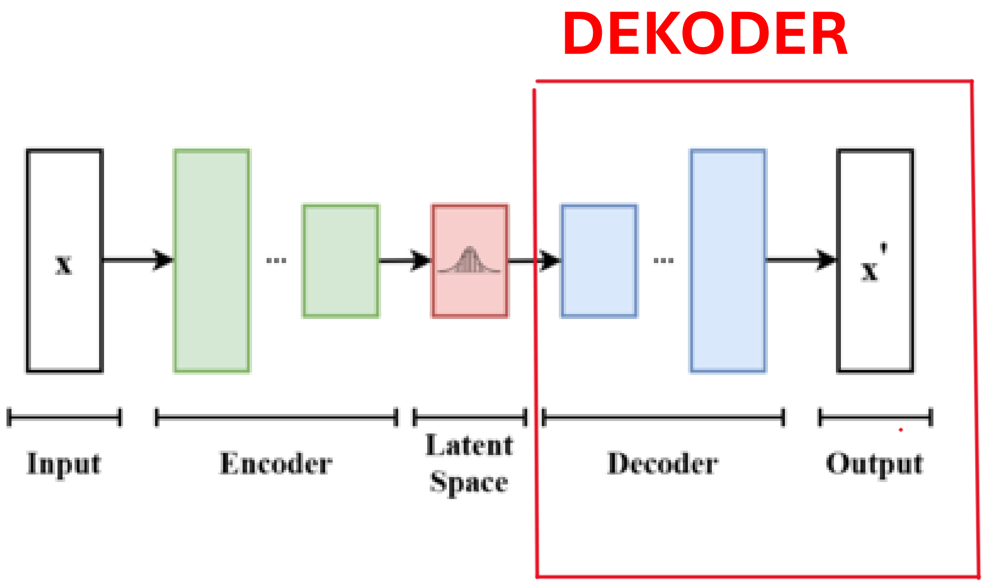
$$L = L_R + L_{KL}$$

Generowanie danych

1. Wygeneruj dane z *d*-wymiarowego rozkładu normalnego, gdzie d to wymiar przestrzeni letalnej.
2. Przekształć dane używając wyuczonych parametrów μ i σ.
3. Oblicz wyjście dekodera.
4. Opcjonalnie wyostrz wyjście



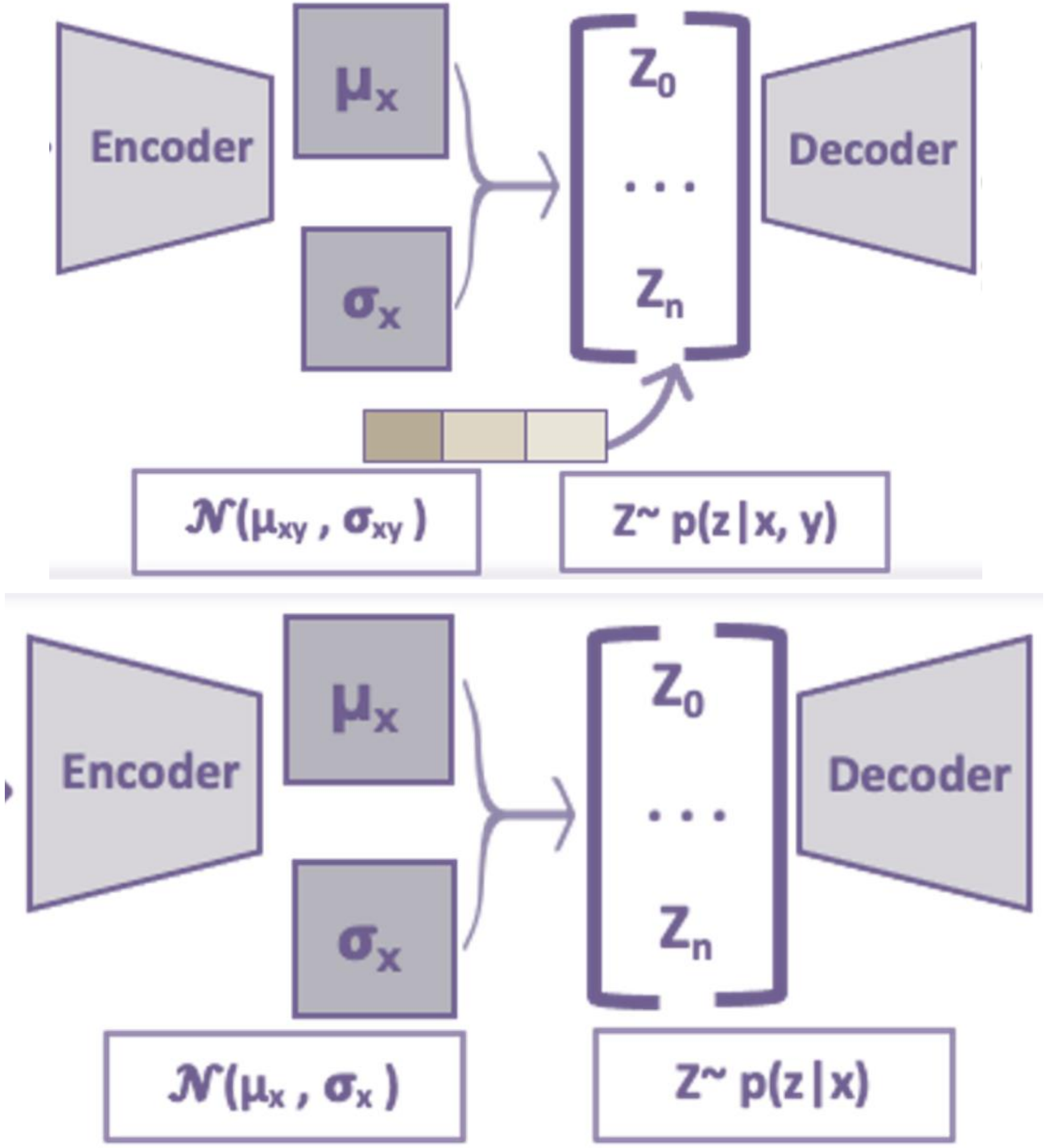
Generowanie danych o określonych cechach



1. Wiemy, że w przestrzeni letalnej znajdują się obiekty o jednej z pożądanych cech w okolicy punktu z_1 (np. niebieskie oczy).
2. Wiemy, że w przestrzeni letalnej znajdują się obiekty o drugiej z pożądanych cech w okolicy punktu z_1 (np. czerwone włosy).
3. Szukamy interesującego nas obiektu między nimi

$$z = (1-\alpha)z_1 + \alpha z_2, \alpha \in [0,1]$$

Conditional Variational Autoencoder, CVAE



Źródło: Sidulova, Mariia, and Chung Hyuk Park. "Conditional variational autoencoder for functional connectivity analysis of autism spectrum disorder functional magnetic resonance imaging data: a comparative study." *Bioengineering* 10.10 (2023): 1209.

Funkcja kosztu

Funkcja straty w VAE składa się z dwóch składników:

1. **Strata rekonstrukcji**

$$L_R = -\mathbb{E}_{q(z|x)} \log p(x|z,y)$$

$$p(x|z,y) = -\frac{1}{2} \|x - DEC(z,y)\|_2^2$$

2. **Dywergencja Kullbacka-Leiblera**

$$L_{KL} = -KL(q(z|x,y)|p(z|y)) = -\int q\,(z|x,y) log \frac{q(z|x,y)}{p(z|y)} dz$$

$$L = L_R + L_{KL}$$

Odległość Wassersteina

Dystans transportu masy

(ang. *earth mover's distance*)

Odległość Wassersteina

Załużmy, że mamy dwie miary (tu probabilistyczne) μ i ϑ na przestrzeni X.

Funkcja kosztu $c(x, y)$ reprezentująca koszt przemieszczenia masy z punktu x w μ do punktu w ϑ.

Jeśli w przestrzeniach zdefiniowanych przez μ i ϑ istnieją momenty rzędu p, to odległość Wassersteina rzędu p jest zdefiniowana jako:

$$W_p(\mu, \vartheta) = \left(\underset{\gamma \in \Pi(\mu, \vartheta)}{inf} \int_{X \times X} c(x, y)^p d\gamma(x, y) \right)^{1/p},$$

gdzie $\Pi(\mu, \vartheta)$ to zbiór wszystkich dopuszczalnych rozkładów sprzężonych γ, takich że:

$$\int_X \; d\gamma(x, y) = \mu(x) \text{ oraz } \int_X \; d\gamma(x, y) = \vartheta(x)$$

Trywialny przykład

Załóżmy, że mamy przestrzeń $X = \{1,2,3,4\}$ oraz miary

$$\mu(x) = \begin{cases} 0.5 & x = 1 \\ 0 & x = 2 \\ 0.5 & x = 3 \\ 0 & x = 4 \end{cases}, \qquad \vartheta(x) = \begin{cases} 0 & x = 1 \\ 0.5 & x = 2 \\ 0 & x = 3 \\ 0.5 & x = 4 \end{cases}$$

Odległość Wassersteina można obliczyć jako minimalny koszt przesunięcia mas, gdzie c$(x, y) = |x - y|$:

Przesunięcie z 1 do 2: 0.5*(2-1) = 0.5

Przesunięcie z 3 do 4: 0.5*(2-1) = 0.5

Całkowity koszt: 0.5+0.5=1.0.

Mniej trywialny przykład

- Rozważmy dwa dwuwymiarowe rozkłady normalne:

1. **Rozkład źródłowy** μ: $\mu = N(\mu_1, \Sigma_1)$, gdzie:
    - Średnia $\mu_1 = (0,0)$,
    - Macierz kowariancji $\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$
2. **Rozkład docelowy** ϑ: $\vartheta = N(\mu_2, \Sigma_2)$, gdzie:
    - Średnia $\mu_2 = (3, 3)$
    - Macierz kowariancji $\Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

Mniej trywialny przykład

Odległość Wassersteina dla rozkładu normalnego:

$$W_1(\mu, \vartheta) = \|\mu_1 - \mu_2\| + tr\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)^{1/2},$$

gdzie $\Sigma_1^{1/2}$ to uzyskana w wyniku dekompozycji Cholesky-ego macierzy Σ_1

Funkcja kosztu

Funkcja straty w WAE składa się z dwóch składników:

1. **Strata rekonstrukcji**

$$L_R = -\mathbb{E}_{q(z|x)} \log p(x|z)$$

$$p(x|z) = -\frac{1}{2} \|x - DEC(z)\|_2^2$$

2. **Odległość Wasserstaina**

$$L_{W_p} = D(q(z), p(z))$$

$$L = L_R + \lambda L_{W_p}$$

Zalety WAE nad VAE

1. **Unikanie problemów z KL-divergence**:
   - WAE nie wymaga warunków nakładających się obszarów wsparcia między q(z) a p(z), dzięki czemu generuje bardziej realistyczne próbki.
2. **Lepiej ukształtowana przestrzeń latentna**:
   - Odległość Wassersteina uwzględnia geometrię przestrzeni, co prowadzi do bardziej spójnych reprezentacji.
3. **Mniej rozmytych próbek**:
   - W porównaniu z VAE, WAE generuje próbki o wyższej jakości, bardziej zbliżone do danych rzeczywistych.

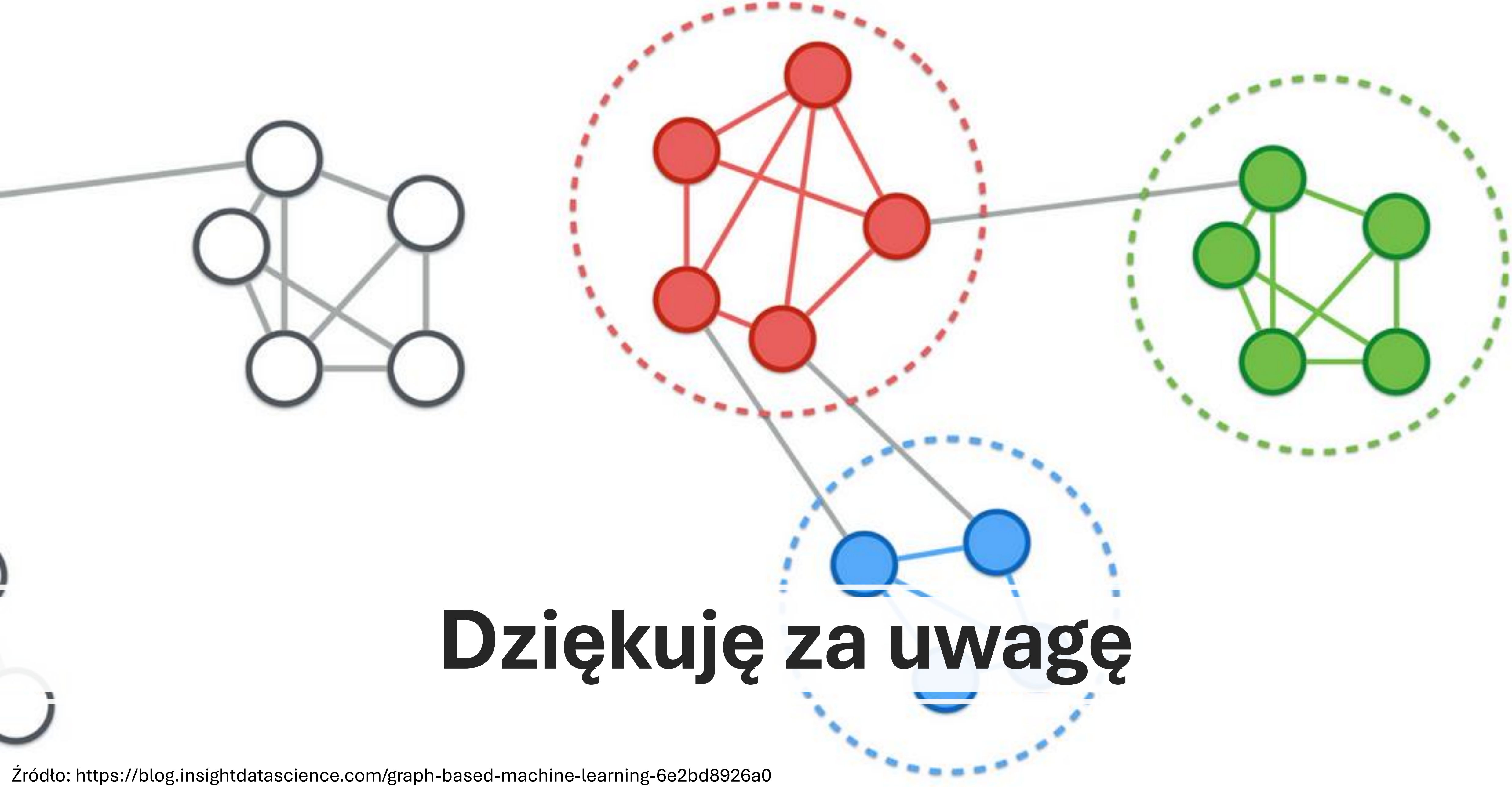
Sinkhorn Distance:

https://amsword.medium.com/a-simple-introduction-on-sinkhorn-distances-d01a4ef4f085

Maximum mean discrepancy

https://www.onurtunali.com/ml/2019/03/08/maximum-mean-discrepancy-in-machine-learning.html

Pytania / Komentarze / Uwagi



Dziękuję za uwagę

Źródło: https://blog.insightdatascience.com/graph-based-machine-learning-6e2bd8926a0