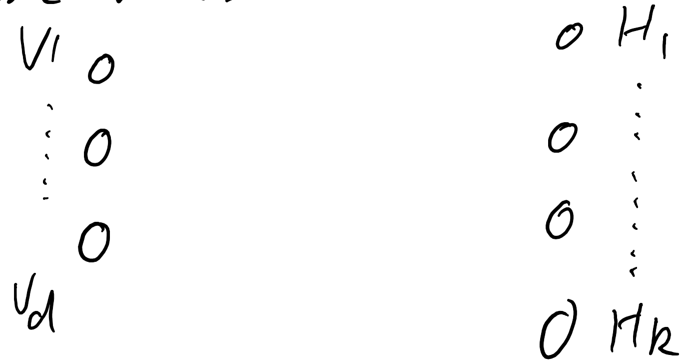
L11 EC500K

→ Unsupervised DL

- Three approaches:
  - ML based:
    DBN, RBM, DBM
  - Reconstruction error:
    Autoencoders
  - Adversarial training:
    GAN

→ RBM:

- Visible units Hidden units



$W \in \mathbb{R}^{d \times k}$

$b \in \mathbb{R}^{1 \times k}$

$a \in \mathbb{R}^{1 \times d}$

- Classical RBM's units are binary,
  $\theta: \{W, b, a\} \Rightarrow$ it has 2^{btk} possibility of v & h
- Energy function, $E(v,h)$:
  $$-\sum_{i=1}^{d} a_i v_i - \sum_{j=1}^{k} b_j h_j - \sum_{i=1}^{d}\sum_{j=1}^{k} w_{ij} v_i h_j$$

$$= -a^T v - b^T h - v^T W h$$

- Joint prob dist:

$$P(v,h;\theta) = \frac{1}{Z} \cdot e^{-E(v,h)}, \quad \text{normalizer: } Z = \sum_{v,h} e^{-E(v,h)}$$

- Marginal P:

$$P(v;\theta) = \frac{1}{Z} \sum_h e^{-E(v,h)}$$

- Cond P:

$$P(h|v;\theta) = \frac{\cancel{\frac{1}{Z}}\, e^{-E(v,h)}}{\cancel{\frac{1}{Z}} \sum_h e^{-E(v,h)}} = \frac{e^{\cancel{a^T v} + b^T h + v^T W h}}{\sum_h e^{\cancel{a^T v} + b^T h + v^T W h}}$$

$$\propto e^{\sum^{k} b_j h_j + \sum^{d} \sum^{k} w_{ij} v_i h_j} = \prod^{k} e^{b_j h_j + (w_j^T v) h_j}$$

$\tilde{p}(h_j | v; \theta)$

$$\Rightarrow P(h_j = 1 | v; \theta) = \frac{\tilde{p}(h_j = 1 | v; \theta)}{\tilde{p}(h_j = 1 | v; \theta) + \tilde{p}(h_j = 0 | v; \theta)}$$

$$= \sigma(w_j^T v + b_j)$$

- Different energy func gives different cond P
- $P(v|h;\theta) = \prod_{i=1}^{d} P(v_i | h; \theta) = \sigma(\vec{h} w_i + a_i)$

you can go backwards with this

- Generative model: since we know joint prob.
- ML based method to learn θ

- $P(v;\theta) = \frac{1}{Z}\sum_h e^{-E(v,h)}$

- observe $v_1 \cdots v_n$:

- $P(V;\theta) = \prod^n P(V_i;\theta) \Rightarrow \log P(V;\theta) =$

$$\sum_{p=1}^n \left(\log\left(\sum_h e^{-E(v,h)}\right) - \log\left(\sum_v \sum_h e^{-E(v,h)}\right)\right)$$

$$\frac{\partial}{\partial w_{ij}} E(v,h) = -v_i h_j$$

$$\frac{\partial}{\partial w_{ij}} \log P(V;\theta) = \sum_{p=1}^n \frac{\sum_h e^{-E(v^{(p)},h)} \cdot v_i h_j}{\sum_h e^{-E(v^{(p)},h)}} - n \sum_{vh} P(v,h)\, v_i h_j$$

$$= \underbrace{\sum_{p=1}^n \sum_h P(h|v^{(p)};\theta)}_{E_{data}} \cdot v_i h_j - n \underbrace{\sum_{v,h} P(v,h)}_{E_{model}} \cdot v_i h_j$$

$$\Rightarrow \log P(V;\theta) = \sum_{p=1}^n \left[E_{data}(v_i^{(p)} h_j) - E_{model}(v_i h_j) \right]$$

Too hard to calculate mean, 2^{d+k} combinations

so we estimate: first term

sample h_j for each $v_i \Rightarrow$ unbiased estimate.

second term:

we need to sample $P(v,h)$: Gibbs sampling

Repeat:

sample $\tilde{h} \sim p(h|v)$

sample $\tilde{v} \sim p(v|h)$

set $v = \tilde{v}$, $h = \tilde{h}$

return $P(v,h)$

- This way of estimate gradiant with sampling is called Contrastive Divergence.