Masked language modeling

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\boxed{M, O \sim \gamma(T_n)}}\left[\sum_{i=1}^{T_m^n} \log p(x_{m_i} \mid x_{o_1}, \dots, x_{o_{T_o}}; \theta)\right].$$

① MLM ⇒ AR-LM

left-to-right ordering

$$p(X) = \prod_{t=1}^{T} p(x_t \mid x_1, \dots, x_{t-1}, x_{t+1} = \langle mask \rangle, \dots, x_T = \langle mask \rangle)$$

② pseudo-likelihood (Besag, 1971)

$$F(x) = PLL(x) = \sum_{t=1}^{T} \log p(x_t \mid x_1 \dots x_{t-1}, x_{t+1} \dots, x_T)$$

$$p(X) = \frac{\exp\{F(x)\}}{\sum_{x' \in L} \exp\{F(x')\}}$$, where L: all possible text snippets.

③ $l \sim \text{length}(\dots)$

$\tilde{X} \sim \text{random}(l)$

for $k = 1, \dots, K$

- uniformly sample an index i at random
- $\tilde{x}_i \sim p(x_i \mid x_1 \dots, x_{i-1}, \langle mask \rangle, x_{i+1} \dots x_T)$

Classification : constrained optimization

$\min_\theta R(\theta)$ ex) $\|\theta\|^2$

subject to

subject to

$d(y_n, f(x_n;\theta)) < \varepsilon$ for all $n = 1, \cdots, N$ — semi-supervised learning

per-example loss.

$d'(f'(x'_m;\theta)) < \delta$ for all $m = 1, \cdots, M$

$$\min_\theta \frac{1}{N} \sum_{n=1}^{N} d(y^n, f(x^n;\theta)) + \lambda R(\theta)$$

$\lambda \geq 0$

SGD

SGD cannot move too far.

Obs 1) SG. : an unbiased estimate of the gradient

$$\left\| \nabla - \mathbb{E}_{B \sim \text{batches}} \left[\sum_{x \in B} \frac{1}{|B|} \nabla_\theta l(x;\theta) \right] \right\|^2 = 0$$

$\nabla \to 0$

Obs 2)

$$\text{Cov}_{B \sim \text{batches}} [\nabla_B] \gg 0$$

Incorporate unlabelled examples for initialization.

stage 1 (pretraining)

$$\theta_0 = \underset{\theta}{\text{(approximate) arg min}} \frac{1}{M} \sum_{m=1}^{M} d'(f'(x'_m;\theta))$$

density modeling

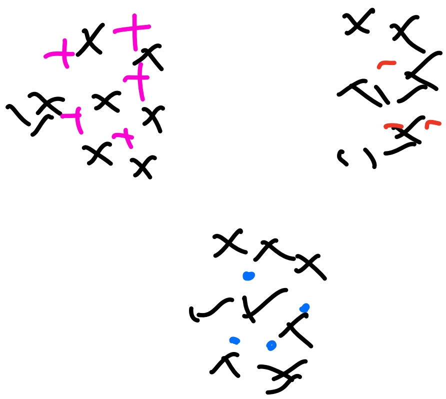
stage 2 (finetuning)

$$\theta_{t+1} = \theta_t - \eta \frac{1}{|B|} \sum_{(x,y) \in B} \nabla_\theta d(y, f(x;\theta))$$

$$p(y|x) = \underbrace{\frac{p(x|y)\,p(y)}{p(x)}}_{\text{Bayes's rule.}} \propto \underbrace{\overbrace{p(x|y)\,p(y)}^{=p(x,y)}}_{\text{Bayes' classifier}} = \underbrace{p(y)\prod_{i=1}^{d} p(x_i|y)}_{\text{Naive Bayes' classifier}}$$

discriminative

Generative



Classification RBM

$$\underbrace{-E(x,y)}_{\text{score}} = b_x^T x + b_y^T y + \sum_{k=1}^{K} \log\left(1 + \exp\{w_k^T x + u_k^T y + c_k\}\right).$$

$$p(x,y) = \frac{\exp\{-E(x,y)\}}{\sum_{y\in C}\int_x \exp\{-E(x,y)\}\,dx}$$

losses

$$p(y|x) = \frac{\exp\{-E(x,y)\}}{\sum_{y'\in C}\exp\{-E(x,y')\}}$$

$$p(x) = \sum_{y'\in C} p(x,y')$$

data

$\underbrace{x'_m}_{\text{unlabelled}}$? $\underbrace{(x_n, y_n)}_{\text{labelled}}$

$\Downarrow$

$$\nabla_\theta \log \sum_y p(x,y) = \mathbb{E}_{y \sim Q(y)} [\nabla_\theta \log p(x,y)]$$

$$Q(y) = \frac{\exp\{-E(x,y)\}}{\sum_{y'} \exp\{-E(x,y')\}} = p(y|x)$$

$$\nabla_\theta \log p_\theta(x,y) = -\nabla_\theta E(x,y) + \underline{\mathbb{E}_{x',y' \sim p_\theta(x,y)} [\nabla E(x',y')]}$$

$$-E(x,y,h) = b_x^T x + b_y^T y + \sum_{k=1}^{K} \underbrace{h_k}_{\{0,1\}} (w_k^T x + u_k^T y + c_k).$$

$$p(h_k = 1 | x,y) = \overset{\text{sigmoid}}{\sigma} (w_k^T x + u_k^T y + c_k)$$

$$p(y = c | h) \propto \exp\{\bar{u}_c^T h + b_{y,c}\}$$

$$p(x_i = 1 | h) = \sigma(\bar{w}_i^T h + b_x^i).$$

$$\bar{U} = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_K^T \end{bmatrix}$$

c^{th} $\bar{u}_c$

Word embeddings : $\underline{\text{continuous}}$ $\underline{\text{b}}$ag-$\underline{\text{o}}$f-$\underline{\text{w}}$ords (CBoW)

- $p(x_{m_1}, \ldots, x_{m_{T_m}} | \text{corrupt}(x))$

$$= \prod_{i=1}^{T_m} p(x_{m_i} | \underline{\text{corrupt}(x)})$$

$$\approx \prod_{i=1}^{T_m} p(x_{m_i} | x_{m_i - n/2}, \ldots, x_{m_i - 1}, \langle \text{mask} \rangle, x_{m_i+1}, \ldots, x_{m_i + n/2}).$$

softmax

$$F(x_{m_i - n/2} \ldots x_{m_i} \ldots x_{m_i + n/2}) = e(x_m)^T \left(\sum_{j=m-n/2}^{m-1} e(x_j) + \sum_{j=m+1}^{m+n/2} e(x_j) \right)$$

CBoW