distributional hypothesis

$\Downarrow$

language modeling

$X = \underbrace{(x_1, x_2, \cdots, x_T)}_{\text{text}}$ $\quad x_t \in V$

$p(X) = p(x_1, x_2, \cdots, x_T)$

① fully factorial : $p(x_1, x_2, \cdots, x_T) = \prod_{t=1}^{T} p(x_t)$

② a chain rule of probability

$p(x_1, x_2 \cdots, x_T) = p(x_{o(1)})\, p(x_{o(2)} | x_{o(1)})\, p(x_{o(3)} | x_{o(1)} x_{o(2)}) \cdots p(x_{o(T)} | \cdots)$

$o : \{1, \cdots, T\} \to \{1, \cdots, T\}$

$o(i) = i$

③ latent variable models.

$p(x_1 \cdots, x_T) = \int p(z) \prod_{t=1}^{T} p(x_t | z)\, dz$

---

the goal of LM.

- $p(X; \theta) > p(x' | \theta)$ if $X \sim D^*$ & $x' \not\sim D^*$

- unsupervised learning.

---

learning. maximum log-likelihood learning

D : data / corpus

$$\mathcal{L}(\theta; D) = -\frac{1}{|D|} \sum_{x \in D} \log p(x; \theta) \qquad -\frac{1}{|D|} \sum_{x,y \in D} \log p(y \mid x; \theta)$$

Autoregressive models

$$\mathcal{L}(\theta; D) = -\frac{1}{|D|} \sum_{x \in D} \log p(x; \theta)$$

$$= \underbrace{-\frac{1}{|D|} \sum_{n=1}^{|D|} \underbrace{\sum_{t=1}^{T^n} \underbrace{\log p(x_{o(t)} \mid x_{o(1):o(t-1)}; \theta)}_{\text{per-step}}}_{\text{per-example}}}_{\text{total}}$$

$$\log p(x_t \mid x_1 \cdots x_{t-1}; \theta)$$

$$p(x_t \mid x_{<t}) = \frac{p(x_{<t}, x_t)}{p(x_{<t})} = \frac{p(x_{<t}, x_t)}{\sum_{x' \in V} p(x_{<t}, x_t = x')}$$

$$= \frac{\tilde{p}(x_{<t}, x_t) / \cancel{Z}}{\sum_{x' \in V} \tilde{p}(x_{<t}, x_t = x') / \cancel{Z}} = \frac{c(x_{<t}, x_t)}{\sum_{x' \in V} c(x_{<t}, x_{t'})}$$

Count-based estimation

N tosses: (M) heads $(N-M)$ tails

$$p(H) = \frac{M}{N} \qquad p(T) = \frac{N-M}{N}.$$

Maximum likelihood estimation
(count-based)

$$p(E_i) = \frac{c(E_i)}{\sum_j c(E_j)}$$

Sparsity.

$$|V|^t \gg |D|$$

0-count

---

n-th order Markov assumption

$$p(x_t | x_1 \cdots, x_{t-1}) = p(x_t | x_{t-n} \cdots x_{t-1})$$

$$\frac{|D|}{O(|V|^n)} \gg \frac{|D|}{O(|V|^t)}$$

$$p(x_t | x_{<t}) = p(x_t | x_{t-n:t-1}) \approx \frac{c(x_{t-n}, \cdots, x_t)}{\sum_{x' \in V} c(x_{t-n}, \cdots, x')}$$

Smoothing: MAP version

$$p(x_t | x_{<t}) = p(x_t | x_{t-n:t-1}) = \frac{c(x_{t-n}, \cdots, x_t) + \varepsilon}{\sum_{x' \in V} c(x_{t-n}, \cdots, x') + |V|\varepsilon}$$

$$d(w, v) = \begin{cases} 0, & \text{if } w = v \\ 1, & \text{otherwise} \end{cases}$$

Feedforward language model

$$F(x_{t-n}, x_{t-n+1}, \cdots, x_{t-1}, x) = u_x^T \phi(x_{t-n}, \cdots, x_{t-1}) + b_x$$