Skip-gram model

Daniel Hsu

February 17, 2025

This note follows Gittens et al (2017) to explain the vector additivity property for analogies, but here using the language of exponential families.

- Skip-gram model (of Mikolov et al) is probability model for words and their context
- Context: finite set of "surrounding" words
- Data point: $(w, c_1, \ldots, c_m) = (w, c_{1:m}) \in \mathcal{W} \times \mathcal{C}^m$ for some m
- Model specification:
  - Distribution of w given single context element $c \in \mathcal{C}$ is a Gibbs distribution

    $$p(w \mid c) := \frac{1}{Z_c} \exp(T(w) \cdot U(c))\pi(w)$$

    where

    $$Z_c := \sum_{w' \in \mathcal{W}} \exp(T(w) \cdot U(c))\pi(w).$$

    * π is a base measure on $\mathcal{W}$
    * $T \colon \mathcal{W} \to \mathbb{R}^d$ and $U \colon \mathcal{C} \to \mathbb{R}^d$ are feature maps (regarded as parameters of the model)
  - Conditional independence property: $c_1, \ldots, c_m$ are conditionally independent given w

    $$p(c_{1:m} \mid w) = \prod_{i=1}^m p(c_i \mid w)$$

- Prescribed parametric form for $p(w \mid c)$ and conditional independence property puts constraints on form of the distribution of $(w, c_1, \ldots, c_m)$
- In particular:

$$\begin{aligned}
p(w \mid c_{1:m}) &= \frac{p(w)p(c_{1:m} \mid w)}{p(c_{1:m})} \quad \text{(Bayes' rule)} \\
&= \frac{p(w)}{p(c_{1:m})} \prod_{i=1}^m p(c_i \mid w) \quad \text{(conditional independence)} \\
&= \frac{p(w)}{p(c_{1:m})} \prod_{i=1}^m \frac{p(c_i)p(w \mid c_i)}{p(w)} \quad \text{(Bayes' rule)} \\
&= \frac{\prod_{i=1}^m p(c_i)}{p(c_{1:m}) \prod_{i=1}^m Z_{c_i}} \exp\left(T(w) \cdot \sum_{i=1}^m U(c_i)\right) p(w) \left(\frac{\pi(w)}{p(w)}\right)^m \quad \text{(parametric form for } p(w \mid c_i))
\end{aligned}$$

- As the context $c_{1:m}$ is varied, the form of $p(w \mid c_{1:m})$ specifies an exponential family of distributions over $\mathcal{W}$
  - Distributions are Gibbs distribution with base measure $p(w)(\pi(w)/p(w))^m$ and feature function $T(w)$
  - Natural parameter corresponding to $c_{1:m}$ is given by $\sum_{i=1}^m U(c_i)$
  - Call this family $\mathcal{Q}_0 := \{q_{c_{1:m}} : c_{1:m} \in \mathcal{C}^m\}$ where $q_{c_{1:m}} = p(\cdot \mid c_{1:m})$
- The parametric form of $p(w \mid c)$ (for single context c) also specifies an exponential family of distributions over $\mathcal{W}$
  - Distributions are Gibbs distribution with base measure $\pi(w)$ and feature function $T(w)$
  - Natural parameter corresponding to single context c is given by $U(c)$
  - Call this family $\mathcal{Q}_1 := \{q_c : c \in \mathcal{C}\}$ where $q_c = p(\cdot \mid c)$
- Definition: single context $c \in \mathcal{C}$ is a "paraphrase" of the context $c_{1:m} \in \mathcal{C}^m$ if $q_c = q_{c_{1:m}}$
- If $\pi(w) \equiv p(w)$, then the two exponential families $\mathcal{Q}_0$ and $\mathcal{Q}_1$ specified above are the same (even for $m > 1$)
  - They have the same feature function, and since $\pi(w) \equiv p(w)$, they have the same base measure
  - In this case, "c is a paraphrase of $c_{1:m}$" is equivalent to

    $$U(c) = \sum_{i=1}^m U(c_i)$$

    because LHS and RHS are two ways to write the natural parameters for the same member of the exponential family
- Vector additivity property for analogies like "man : woman :: king : queen":
  - Hypothesis: there is a common "relation" R (either a single context or multiple context elements) such that
    1. "man" is a paraphrase of ("woman", R)
    2. "king" is a paraphrase of ("queen", R)
  - Under the assumtion $\pi(w) = p(w)$, we must have

    $$\begin{aligned} U(\text{“man”}) &= U(\text{“woman”}) + U(R) \\ U(\text{“king”}) &= U(\text{“queen”}) + U(R) \end{aligned}$$

    which implies

    $$\begin{aligned} U(\text{“queen”}) &= U(\text{“king”}) - U(R) \\ &= U(\text{“king”}) - U(\text{“man”}) + U(\text{“woman”}) \end{aligned}$$

The first "theorem" of Gittens et al (2017) concerns minimizers of a relative entropy $\mathrm{RE}(q, p_\lambda)$ over a family of Gibbs distributions $\mathcal{P} = \{p_\lambda : \lambda \in \mathbb{R}^d\}$ corresponding to feature map $T \colon \mathcal{W} \to \mathbb{R}^d$ and base measure π. Recall that $p_\lambda(w) = \exp(\lambda \cdot T - G(\lambda))\pi(w)$, where $G(\lambda) = \log \pi[\exp(\lambda \cdot T)]$. (It is not important what q is.) We can write

$$\mathrm{RE}(q, p_\lambda) = q\left[\log \frac{q}{p_\lambda}\right] = -\mathrm{H}(q) - q[\log p_\lambda] = -\mathrm{H}(q) - q[\lambda \cdot T - G(\lambda) + \log \pi],$$

so the gradient with respect to λ is

$$-q[T] + \nabla G(\lambda).$$

Recall that the gradient of the log partition function $G(\lambda)$ is precisely $p_\lambda[T]$, and hence the gradient of $\mathrm{RE}(q, p_\lambda)$ is zero precisely when $q[T] = p_\lambda[T]$.