Pattern Recognition Letters
journal homepage: www.elsevier.com

Supplementary Material - Boosted Negative Sampling by Quadratically Constrained Entropy Maximization

Taygun Kekeç[a], David Mimno[b], David M. J. Tax[a]

[a] *Pattern Recognition and Bioinformatics Laboratory*
Delft University of Technology
Mekelweg 4 2628CD, Delft, The Netherlands
[b] *Information Sciences Department, Cornell University*
Ithaca, NY 14853, New York

ABSTRACT

This is the supplementary material for the paper titled "Boosted Negative Sampling by Quadratically Constrained Entropy Maximization".

Appendix A. Negative Sampling Objective

The negative sampling objective is given as follows:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{p_d}\left[\ln \sigma(\boldsymbol{x};\boldsymbol{\theta})\right] + \mathbb{E}_{p_n^0}\left[\ln(1-\sigma(\boldsymbol{y};\boldsymbol{\theta}))\right] \tag{A.1}$$

Here, σ is the sigmoid function:

$$\sigma(\boldsymbol{u};\boldsymbol{\theta}) = \frac{1}{1+\exp\left[-G(\boldsymbol{u};\boldsymbol{\theta})\right]}$$

where G is the difference between the log likelihood of the sample under the model and the negative sampling distribution:

$$G(\boldsymbol{u};\boldsymbol{\theta}) = \ln p_m^{\boldsymbol{\theta}}(\boldsymbol{u}) - \ln p_n(\boldsymbol{u})$$

Substitution of σ and G functions gives us the following:

$$J_T(\boldsymbol{\theta}) = E_{p_d}\left[\ln \frac{p_m^{\boldsymbol{\theta}}(\boldsymbol{x})}{p_m^{\boldsymbol{\theta}}(\boldsymbol{x}) + p_n(\boldsymbol{x})}\right] + \mathbb{E}_{p_n^0}\left[\ln \frac{p_n(\boldsymbol{y})}{p_m^{\boldsymbol{\theta}}(\boldsymbol{y}) + p_n(\boldsymbol{y})}\right]$$

Using logarithmic properties and expectation additivity, we decompose this objective into:

$$\begin{aligned} J(\boldsymbol{\theta}, p_n) = {} & \mathbb{E}_{p_d}[\ln p_m^{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{p_d}[\ln(p_m^{\boldsymbol{\theta}}(\boldsymbol{x}) + p_n(\boldsymbol{x}))] \\ & -\mathbb{E}_{p_n^0(y)}[\ln(p_m^{\boldsymbol{\theta}}(\boldsymbol{y})) + p_n(\boldsymbol{y}))] + \mathbb{E}_{p_n^0(y)}[\ln p_n(\boldsymbol{y})] \end{aligned}$$

where fourth term is constant in $\boldsymbol{\theta}$. □

Appendix B. Smoothing the distribution

Assume we have a probability mass function, with *ordered* entries:

$$p_1 \geq p_2 \geq \ldots \geq p_n > 0 \tag{B.1}$$

with $\sum_{i=1}^{n} p_i = 1$. We smooth PMF p slightly, by modifying two neighbouring probabilities with a small probability Δ_i. This defines a new PMF $\tilde{p}$, with $\tilde{p}_i = p_i - \Delta_i$, $\tilde{p}_{i+1} = p_{i+1} + \Delta_i$, and all other probabilities remain the same. The entropy change: $H(\tilde{p}) - H(p)$ can be stated as:

$$\begin{aligned} = {} & -(p_i - \Delta_i)\log(p_i - \Delta_i) - (p_{i+1} + \Delta_i)\log(p_{i+1} + \Delta_i) \\ & + p_i \log p_i + p_{i+1}\log p_{i+1} \\ = {} & -p_i(\log(p_i - \Delta_i) - \log p_i) - p_{i+1}(\log(p_{i+1} + \Delta_i) - \log p_{i+1}) \\ & + \Delta_1 \log(p_i - \Delta_i) - \Delta_i \log(p_{i+1} + \Delta_i) \\ = {} & -p_i(\log(1 - \frac{\Delta_i}{p_i})) - p_{i+1}(\log(1 + \frac{\Delta_i}{p_{i+1}})) \\ & + \Delta_i \log(p_i(1 - \frac{\Delta_i}{p_i})) - \Delta_i \log(p_{i+1}(1 + \frac{\Delta_i}{p_{i+1}})) \end{aligned}$$

The logarithms are of the form $\log(1 + x)$ for which the Taylor expansion around $x = 0$ can be used:

$$\log(1 + x) = 0 + x + O(x^2) \tag{B.2}$$

e-mail: taygunkekec@gmail.com (Taygun Kekeç), mimno@cornell.edu (David Mimno), D.M.J.Tax@tudelft.nl (David M. J. Tax)

Therefore, the substitution gives:

$$
\begin{aligned}
H(\tilde{p}) - H(p) &= p_i \frac{\Delta_i}{p_i} - p_{i+1}\frac{\Delta_i}{p_{i+1}} \\
&\quad +\Delta_i \log p_i - \Delta_i \frac{\Delta_i}{p_i} - \Delta_i \log p_{i+1} - \Delta_i \frac{\Delta_i}{p_{i+1}} + O(\Delta_i^2) \\
&= +\Delta_i \log p_i - \Delta_i \log p_{i+1} + O(\Delta_i^2) \\
&= \Delta_i \log \frac{p_i}{p_{i+1}} + O(\Delta_i^2) > 0
\end{aligned} \tag{B.3}
$$

The first two terms cancel, the forth and the sixth are of order $O(\Delta_i^2)$, and only the third and fifth term remain. Because $p_i > p_{i+1}$, this difference between the entropies $H(\tilde{p}) - H(p)$ is larger than 0. □

Appendix C. Powering the distribution

Assuming we have a probability mass function, as defined in Equation (B.1). We define a power λ, $0 < \lambda < 1$, and rescale the PMF:

$$\tilde{p}_i = \frac{p_i^\lambda}{\sum_j p_j^\lambda} \tag{C.1}$$

This new distribution is more smooth when

$$\hat{\Delta}_i \leq \Delta_i \tag{C.2}$$

where $\Delta_i = p_i - p_{i+1}$ That would mean:

$$
\begin{aligned}
\frac{p_i^\lambda - p_{i+1}^\lambda}{\sum_j p_j^\lambda} &\leq p_i - p_{i+1} \\
p_i^\lambda - p_{i+1}^\lambda &\leq (\sum_j p_j^\lambda)(p_i - p_{i+1}) && \text{(C.3)} \\
p_i^\lambda - p_{i+1}^\lambda &\leq C(p_i - p_{i+1}) && \text{(C.4)}
\end{aligned}
$$

This is actually the definition of Lipschitz continuity (Mohri et al., 2012). Unfortunately, for $f(x) = x^\lambda$ where $x \in [0, 1]$ and $0 < \lambda < 1$ function f is *not* Lipschitz continuous, because for very small values of x the derivative goes to infinity.

If we now assume that $\gamma < p_i < 1$, our purpose is to derive a lower bound γ for p_i such that (C.3) actually holds. First, we define the function f:

$$
\begin{aligned}
f(x) &= = x^\lambda \quad x \in (0, 1),\ \gamma < \lambda < 1 \\
f'(x) &= \lambda x^{\lambda-1} \quad \text{is always positive} \\
f''(x) &= \lambda(\lambda - 1)x^{\lambda-2} \quad \text{is always negative}
\end{aligned}
$$

in other words: the derivative is always positive, but each derivative becomes smaller and smaller. Because we have that $x > \gamma$, and for $h > 0$:

$$f'(\gamma) > f'(x) = \lim_{h\downarrow} \frac{f(x+h) - f(x)}{(x+h) - x} > \frac{f(x+h) - f(x)}{(x+h) - x} \tag{C.5}$$

Using $f'(x) = \lambda x^{\lambda-1}$, and rewriting gives:

$$f(x+h) - f(x) < \lambda\gamma^{\lambda-1}((x+h) - x) \tag{C.6}$$

Substitution of $x + h = p_i$ and $x = p_{i+1}$ and solving γ reads:

$$p_i^\lambda - p_{i+1}^\lambda < \lambda\gamma^{\lambda-1}(p_i - p_{i+1}) \tag{C.7}$$

Now we can identify γ using Equation (C.3):

$$
\begin{aligned}
\lambda\gamma^{\lambda-1} &= \sum_j p_j^\lambda && \text{(C.8)} \\
\gamma &= \left(\frac{1}{\lambda}\sum_j p_j^\lambda\right)^{1/(\lambda-1)} && \text{(C.9)}
\end{aligned}
$$

Now, γ gamma is lower bounded as such, powering the distribution acts as a smoother. □

References

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of Machine Learning. The MIT Press.