Noise Contrastive Estimation

Karl Stratos

In prediction problems, we're supposed to predict $y \in \mathcal{Y}$ from $x \in \mathcal{X}$. We do this by assuming a joint distribution p_{XY} from which we can sample correct pairs (x, y) and learning a score function $s^\theta(x, y) \in \mathbb{R}$ parameterized by θ such that it assigns a high score to a correct pair and a low score to an incorrect pair. To estimate such a score function, we often use the hinge loss (Appendix A) or the cross-entropy loss (Appendix B)

In **noise constrastive estimation (NCE)**, we choose a "noise" distribution q_Y over $\mathcal{Y}$ and the size of a sample set N and consider the task of distinguishing true samples from fake samples. It underlies many successful methods such as the skip-gram model [5], the generative adversarial networks (GANs) [1], and contrastive predictive coding [6]. It has two popular formulations: 1. **global**: infer which of the N samples is true, and 2. **local**: for each individual sample infer if it's true.

Information theory enables a simple and insightful analysis of NCE. Given any distribution p, if q^θ is a distribution over the same variables parameterized by θ, q^θ is equal to p iff it is the minimizer of the cross entropy between p and q^θ

$$\theta^* \in \arg\min_\theta \mathop{\mathbf{E}}_{z \sim p}\left[-\ln q^\theta(z)\right] \qquad \Longleftrightarrow \qquad q^{\theta^*}(z) = p(z) \qquad \forall z$$

assuming the **universality** of q^θ: that is, it is expressive enough to model p. While universality should be assumed with a grain of salt (e.g., it might require an exponentially large parameter space), it seems to hold in practice with neural networks and greatly simplifies the analysis.

1 Global NCE

1.1 Model

The global NCE objective assumes a joint distribution

$$p_{IXY^N}(i, x, y_1 \ldots y_N) := \frac{1}{N} p_{XY}(x, y_i) \prod_{j \neq i} q_Y(y_j)$$

That is, we first draw an index $i \in \{1 \ldots N\}$ *uniformly* at random and for $j = 1 \ldots N$ draw $(x, y_j) \sim p_{XY}$ if $j = i$ but otherwise draw $y_j \sim q_Y$. This yields a conditional distribution over indices

$$p_{I|XY^N}(i|x, y_1 \ldots y_N) = \frac{p_{Y|X}(y_i|x) \prod_{j \neq i} q_Y(y_j)}{\sum_{k=1}^N p_{Y|X}(y_k|x) \prod_{j \neq k} q_Y(y_j)} = \frac{\frac{p_{Y|X}(y_i|x)}{q_Y(y_i)}}{\sum_{k=1}^N \frac{p_{Y|X}(y_k|x)}{q_Y(y_k)}} \tag{1}$$

Let $H(I|XY^N)$ denote the entropy of $p_{I|XY^N}$. The following observation is made in [6].

Lemma 1.1. *Choose $q_Y = p_Y$. Then $H(I|XY^N) \geq \ln N - I(X,Y)$ where $I(X,Y)$ is the mutual information between $(x,y) \sim p_{XY}$.*

Proof. By (1) and using $q_Y = p_Y$,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(i,x,y_1\dots y_N)\sim p_{IXY^N}} \left[-\ln p_{I|XY^N}(i|x,y_1\dots y_N)\right] \\
&= -\underbrace{\mathop{\mathbf{E}}_{(x,y)\sim p_{XY}} \left[\frac{p_{Y|X}(y|x)}{p_Y(y)}\right]}_{I(X,Y)} + \underbrace{\mathop{\mathbf{E}}_{(i,x,y_1\dots y_N)\sim p_{IXY^N}} \left[\ln \sum_{k=1}^{N} \frac{p_{Y|X}(y_k|x)}{p_Y(y_k)}\right]}_{\gtrapprox \ln N}
\end{aligned}$$

The claim about the second term is nonrigorous but intuitive since conditioning cannot decrease information. $\square$

1.2 Estimation

We use a score function $s^\theta(x,y)$ through the softmax function to parameterize $p_{I|XY^N}$

$$p^\theta_{I|XY^N}(i|x,y_1\dots y_N) := \frac{\exp\left(s^\theta(x,y_i)\right)}{\sum_{j=1}^N \exp\left(s^\theta(x,y_j)\right)} \qquad \forall i \in \{1\dots N\}$$

The model is estimated by minimizing the cross entropy between $p_{I|XY^N}$ and $p^\theta_{I|XY^N}$[1]

$$\theta^* \in \arg\min_\theta \mathop{\mathbf{E}}_{(i,x,y_1\dots y_N)\sim p_{IXY^N}} \left[-\ln \frac{\exp\left(s^\theta(x,y_i)\right)}{\sum_{j=1}^N \exp\left(s^\theta(x,y_j)\right)}\right] \tag{2}$$

By universality we must have $p^{\theta^*}_{I|XY^N} = p_{I|XY^N}$. By (1) this means

$$s^{\theta^*}(x,y) = \ln \frac{p_{Y|X}(y|x)}{q_Y(y)} + \ln C_x \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

for some constant $C_x > 0$. In particular, we can use the optimal parameter θ^* to recover the underlying conditional distribution

$$p_{Y|X}(y|x) = \frac{\exp\left(s^{\theta^*}(x,y) + \ln q_Y(y)\right)}{\sum_{y'} \exp\left(s^{\theta^*}(x,y') + \ln q_Y(y')\right)} \tag{3}$$

This is consistent with the "ranking" algorithm in [3].

[1] Lemma 1.1 implies that if $q_Y = p_Y$ minimizing this objective corresponds to maximizing a lower bound on $I(X,Y)$ that cannot be greater than $\ln N$, which is consistent with the result in [4].

2 Local NCE

2.1 Model

The local NCE objective assumes a biased coin with head probability $1/N$. Given $x \in \mathcal{X}$ and the outcome of a coin flip $a \in \{0,1\}$, it defines

$$p_{Y|XA}(y|x,a) := \begin{cases} p_{Y|X}(y|x) & \text{if } a = 1 \\ q_Y(y) & \text{if } a = 0 \end{cases}$$

This yields the conditional head probability

$$p_{A|XY}(1|x,y) = \frac{p_{Y|X}(y|x)}{p_{Y|X}(y|x) + (N-1)q_Y(y)} \tag{4}$$

Given $x \sim p_X$ and N iid samples $a_i \sim p_A$ and $y_i \sim p_{Y|XA}(\cdot|x, a_i)$ for $i = 1 \dots N$, the joint conditional probability of the coin flips is given by

$$p_{A^N|XY^N}(a_1 \dots a_N | x, y_1 \dots y_N) = \prod_{i=1:a_i=1}^{N} p_{A|XY}(1|x,y_i) \prod_{j=1:a_j=0}^{N} (1 - p_{A|XY}(1|x,y_j))$$

Let $H(A^N|XY^N)$ denote the entropy of $p_{A^N|XY^N}$. Note that we can write it in the familiar form

$$H(A^N|XY^N) = \mathop{\mathbf{E}}_{(x,y)\sim p_{XY}} \left[-\ln p_{A|XY}(1|x,y) \right] + (N-1) \mathop{\mathbf{E}}_{\substack{x \sim p_X \\ y \sim q_Y}} \left[-\ln(1 - p_{A|XY}(1|x,y)) \right] \tag{5}$$

The following lemma can be easily shown by plugging in (4) into (5).

Lemma 2.1. *Let* $\mathrm{KL}(p||q)$ *denote the KL divergence between distributions* p *and* q. *Then*

$$H(A^N|XY^N) = \mathrm{KL}\left(p_{Y|X} \middle\| \frac{p_{Y|X} + (N-1)q_Y}{N} \right) + (N-1)\mathrm{KL}\left(q_Y \middle\| \frac{p_{Y|X} + (N-1)q_Y}{N} \right) + \ln N + \ln\left(\frac{N}{N-1}\right)$$

Corollary 2.2. *Let* $\mathrm{JSD}(p||q)$ *denote the Jensen-Shannon divergence. Then*

$$H(A^2|XY^2) = 2\mathrm{JSD}\left(p_{Y|X} \middle\| \frac{p_{Y|X} + q_Y}{2} \right) + \ln 4 \tag{6}$$

Equation (6) is the GAN objective [1] where q_Y is used as a fixed "generator" and $p_{A|XY}$ is used as an optimal "discriminator" for that generator.

2.2 Estimation

We use a score function $s^\theta(x,y)$ through the sigmoid function to parameterize $p_{A|XY}$

$$p^\theta_{A|XY}(1|x,y) := \frac{1}{1 + \exp\left(-s^\theta(x,y)\right)}$$

This is used to define the joint conditional distribution

$$p^{\theta}_{A^N|XY^N}(a_1 \dots a_N|x, y_1 \dots y_N) = \prod_{i=1:a_i=1}^{N} p^{\theta}_{A|XY}(1|x, y_i) \prod_{j=1:a_j=0}^{N} (1 - p^{\theta}_{A|XY}(1|x, y_j))$$

The model is again estimated by minimizing the cross entropy between $p_{A^N|XY^N}$ and $p^{\theta}_{A^N|XY^N}$. Similar to (5) this objective can be written in the familiar form

$$\theta^* \in \arg\max_{\theta} \mathop{\mathbf{E}}_{(x,y)\sim p_{XY}} \left[\ln p^{\theta}_{A|XY}(1|x, y)\right] + (N-1) \mathop{\mathbf{E}}_{\substack{x\sim p_X \\ y\sim q_Y}} \left[\ln(1 - p^{\theta}_{A|XY}(1|x, y))\right]$$

By universality we must have $p^{\theta^*}_{A|XY} = p_{A|XY}$. By (4) this means

$$s^{\theta^*}(x, y) = \ln \frac{p_{Y|X}(y|x)}{q_Y(y)} - \ln(N-1) \qquad\qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

Note that if $q_Y = p_Y$, the optimal score of (x, y) is the pointwise mutual information (PMI) minus the log of the number of negative examples: this gives the analysis of the skip-gram objective in [2]. We can use the optimal parameter θ^* to recover the underlying conditional distribution

$$p_{Y|X}(y|x) = \exp\left(s^{\theta^*}(x, y) + \ln q_Y(y) + \ln(N-1)\right)$$

This is consistent with the “binary” algorithm in [3]. Note that unlike (3) this calculation doesn’t require normalization. This implies that the score function must self-normalized (Assumption 2.2 in [3]), that is we must be able to at least find θ such that

$$\sum_y \exp\left(s^{\theta}(x, y) + \ln q_Y(y) + \ln(N-1)\right) = 1 \qquad\qquad \forall x \in \mathcal{X}$$

This is a strong assumption when $|\mathcal{X}|$ is larger than the number of variables in θ, so universality cannot be taken for granted in this case.

A Hinge Loss

We want to find θ that maximizes the probability of the event that $s^\theta(x,y) > s^\theta(x,y')$ for all $y' \neq y$. This is equivalent to minimizing the **zero-one loss**

$$\arg\min_\theta \mathop{\mathbf{E}}_{(x,y)\sim p_{XY}} \left[\overbrace{\mathbb{1}\left(\underbrace{s^\theta(x,y) - \max_{y'\neq y} s^\theta(x,y')}_{\text{margin of } (x,y)} \leq 0 \right)}^{\text{zero-one loss on } (x,y)} \right]$$

where $\mathbb{1}(\cdot) \in \{0,1\}$ is the indicator function. The indicator function is difficult to optimize for a number of reasons (e.g., it has zero gradient almost everywhere wrt the margin), so we instead define the **hinge loss**

$$\arg\min_\theta \mathop{\mathbf{E}}_{(x,y)\sim p_{XY}} \left[\overbrace{\max\left\{ 0, 1 - \underbrace{\left(s^\theta(x,y) - \max_{y'\neq y} s^\theta(x,y') \right)}_{\text{margin of } (x,y)} \right\}}^{\text{hinge loss on } (x,y)} \right]$$

Note that for any fixed (x,y), the hinge loss on (x,y) is a convex upper bound on the zero-one loss on (x,y) where the convexity is wrt the margin of (x,y).

In some applications, it's neither necessary nor useful to exactly maximize over the negative space $\{y' \in \mathcal{Y} : y' \neq y\}$ to compute the margin. This is because the search is intractable and/or exact maximization has some undesirable quality (e.g., it's in fact an alternative viable prediction). In this case, maximization is replaced by sampling [7].

B Cross-Entropy Loss

We frame the problem as conditional density estimation of $p_{Y|X}$. To this end, we turn the score function into a proper conditional distribution by using the softmax operation:

$$p^\theta_{Y|X}(y|x) := \frac{\exp\left(s^\theta(x,y)\right)}{\sum_{y'} \exp\left(s^\theta(x,y')\right)} \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

Then we find θ that minimizes the cross entropy between $p_{Y|X}$ and $p^{\theta^*}_{Y|X}$:

$$\theta^* \in \arg\min_\theta \mathop{\mathbf{E}}_{(x,y)\sim p_{XY}} \left[-\ln p^\theta_{Y|X}(y|x) \right]$$

By universality we must have $p^{\theta^*}_{Y|X} = p_{Y|X}$. This means

$$\frac{\exp\left(s^{\theta^*}(x,y)\right)}{\sum_{y'} \exp\left(s^{\theta^*}(x,y')\right)} = \frac{p_{XY}(x,y)}{\sum_{y'} p_{XY}(x,y')} \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

and it follows that $\exp\left(s^{\theta^*}(x,y)\right) = C_x p_{XY}(x,y)$ for some $C_x > 0$. Hence

$$s^{\theta^*}(x,y) = \ln p_{XY}(x,y) + \ln C_x \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

That is, the optimal score of (x,y) is the log probability of (x,y) shifted by some constant dependent on x.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[2] Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

[3] Ma, Z. and Collins, M. (2018). Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv preprint arXiv:1809.01812*.

[4] McAllester, D. and Stratos, K. (2018). Formal limitations on the measurement of mutual information. *arXiv preprint arXiv:1811.04251*.

[5] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[6] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[7] Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.