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 \( \theta \) 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 contrastive 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 \( q^\theta \) is a distribution over the same variables parameterized by \( \theta \), \( q^\theta \) is equal to \( p \) iff it is the minimizer of the cross entropy between \( p \) and \( q^\theta \)

\[
\theta^* \in \arg \min_{\theta} \mathbb{E}_{z \sim p} \left[ -\ln q^\theta(z) \right] \iff q^{\theta^*}(z) = p(z) \quad \forall z
\]

assuming the universality of \( q^\theta \): 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

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

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

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

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 $q_Y = p_Y$,

$$
E_{(i,x,y_1\ldots y_N) \sim p_{I|XY^N}} \left[-\ln p_{I|XY^N}(i|x,y_1\ldots y_N)\right] \\
= -E_{(x,y) \sim p_{XY}} \left[ \frac{p_{Y|X}(y|x)}{p_Y(y)} \right] + \frac{E_{(i,x,y_1\ldots y_N) \sim p_{I|XY^N}} \left[ \ln \sum_{k=1}^{N} p_{Y|X}(y_k|x) \right]}{\leq \ln N}
$$

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 $p_{I|XY^N}$

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

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

$$
\theta^* \in \arg \min_{\theta} E_{(i,x,y_1\ldots y_N) \sim p_{I|XY^N}} \left[-\ln \frac{\exp(s^\theta(x,y_i))}{\sum_{j=1}^{N} \exp(s^\theta(x,y_j))}\right]
$$

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

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

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

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

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

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

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

This yields the conditional head probability

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

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

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

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

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

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

Lemma 2.1. Let $KL(p||q)$ denote the KL divergence between distributions $p$ and $q$. Then

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

Corollary 2.2. Let $JSD(p||q)$ denote the Jensen-Shannon divergence. Then

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

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 $s^0(x,y)$ through the sigmoid function to parameterize $p_{A|XY}$

$$p_{A|XY}^0(1|x,y) := \frac{1}{1 + \exp(-s^0(x,y))}$$

3
This is used to define the joint conditional distribution

\[ p_{A^N|XY^N}(a_1 \ldots a_N|x, y_1 \ldots y_N) = \prod_{i=1}^{N} p_{A|XY}(1|x, y_i) \prod_{j=1}^{N} (1 - p_{A|XY}(1|x, y_j)) \]

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

\[ \theta^* \in \arg \max_\theta E_{(x,y) \sim p_{XY}} \left[ \ln p_{A|XY}^\theta(1|x, y) \right] + (N - 1) E_{y \sim q_Y} \left[ \ln(1 - p_{A|XY}^\theta(1|x, y)) \right] \]

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

\[ s_{\theta^*}(x, y) = \ln \frac{p_{Y|X}(y|x)}{q_Y(y)} - \ln(N - 1) \quad \forall x \in \mathcal{X}, \ y \in \mathcal{Y} \]

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 \( \theta^* \) 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 \( \theta \) such that

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

This is a strong assumption when \(|\mathcal{X}|\) is larger than the number of variables in \( \theta \), so universality cannot be taken for granted in this case.
A Hinge Loss

We want to find $\theta$ 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

$$\arg \min_{\theta} \mathbb{E}_{(x, y) \sim p_{XY}} \mathbb{I} \left( s^\theta(x, y) - \max_{y' \neq y} s^\theta(x, y') \leq 0 \right)$$

where $\mathbb{I}() \in \{0, 1\}$ 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

$$\arg \min_{\theta} \mathbb{E}_{(x, y) \sim p_{XY}} \max \left\{ 0, 1 - \left( s^\theta(x, y) - \max_{y' \neq y} s^\theta(x, y') \right) \right\}$$

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:

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

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

$$\theta^* \in \arg \min_{\theta} \mathbb{E}_{(x, y) \sim p_{XY}} \left[ -\ln p_{Y|X}^\theta(y|x) \right]$$
By universality we must have $p_{Y|X}^\theta = p_{Y|X}$. This means

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

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

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

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

References


