Copula Normalization

Erik Learned-Miller

1 Introduction

The basic idea with copula normalization is a variation on the idea of batch normalization. Instead of normalizing the output distribution of a layer to have mean 0 and variance 1, we instead remap the distribution to be approximately uniform.

Let X be the random variable representing one of the scalar values output by a hidden layer in a network, and suppose that X is distributed according to $f(x)$, an unknown probability density.

If we had access to $f(x)$, then we could compute the cumulative distribution function $F(x)$. We can then form the new random variable

$$U = F(X),$$

through the *probability integral transform*, where U is now uniform.

We can then remap any sample value x_i via $u_i = F(x_i)$ to obtain the normalized value of x_i, and these new values are now distributed uniformly.

In practice, we do not have access to $f(x)$ or $F(x)$, and instead must estimate it from data. We will do this by estimating $M + 1$ specific values on the *inverse cdf* (known as *percentiles*):

$$\{F^{-1}(0), F^{-1}(\frac{1}{M}), F^{-1}(\frac{2}{M}), ..., F^{-1}(\frac{M-1}{M}), F^{-1}(1)\}$$

.

We will use the estimates $\{\hat{F}^{-1}(0), ..., \hat{F}^{-1}(1)\}$ to remap the values, as shown in detail below.

1.1 Mini-batch statistics

Like in the BatchNorm algorithm, we start by estimating some statistics. First consider the percentiles $F^{-1}(\frac{1}{M}), F^{-1}(\frac{2}{M}), ..., F^{-1}(\frac{M-1}{M})$.

Our goal is to estimate the percentile $F^{-1}(\frac{k}{M})$ using a mini-batch sample of size m: $x_1, x_2, ..., x_m$. We will assume that $M << m$.

- Sort the mini-batch samples to form the *order statistics* of the sample $z_1, z_2, ..., z_m$, where z_1 is the smallest sample value.

- If $\frac{k}{M} == \frac{i}{m+1}$ for some integer value i, then $\hat{F}^{-1}(\frac{k}{M}) = z_i$. For example, to estimate $F^{-1}(\frac{3}{8})$ from a sample of size 15, we simply choose $\hat{F}(\frac{3}{8}) = z_6$, since $\frac{3}{8} = \frac{6}{16}$.
- If $\frac{k}{M} == \frac{i+d}{m+1}$ for some integer value i and some $0 < d < 1$, then we must interpolate between z_i and z_{i+1} according to

$$\hat{F}^{-1}(\frac{k}{M}) = (1-d) \cdot z_i + d \cdot z_{i+1}.$$

This strategy works for all of the percentiles except $F^{-1}(0)$ and $F^{-1}(1)$ representing the support of the unknown distribution. In general, it is impossible to say much about the support of a distribution from a set of samples, since distributions can have infinite or finite tails, and in many cases it is impossible to know which.

Therefore, we will use an *extrapolation* method to estimate $F^{-1}(0)$ and $F^{-1}(1)$. The idea is to assume that the unknown distribution is constant over a very small region, and to compute $\hat{F}^{-1}(0)$ as

$$\hat{F}^{-1}(0) = z_0 - (z_1 - z_0) = 2z_0 - z_1.$$

Similarly, we will estimate $\hat{F}^{-1}(1)$ as

$$\hat{F}^{-1}(1) = z_{m-1} + (z_{m-1} - z_{m-2}) = 2z_{m-1} - z_{m-2}.$$

1.2 Normalization using estimated percentiles

The estimated values of the inverse cdf do not completely define a cdf function–there are many values that have not been estimated. However, we will assume that the cdf is *piecewise linear* between the estimated inverse cdf points. Thus, we can produce an estimated cdf from our estimated inverse cdf points via the following formula:

- For $x_i < \hat{F}^{-1}(0)$, $\hat{x}_i \leftarrow 0 - 0.5$.
- For $\hat{F}^{-1}(\frac{k}{M}) \leq x_i < \hat{F}^{-1}(\frac{k+1}{M})$,
  $\hat{x}_i \leftarrow \frac{1}{M+1}\left[k + \frac{x_i - \hat{F}^{-1}(\frac{k}{M})}{\hat{F}^{-1}(\frac{k+1}{M}) - \hat{F}^{-1}(\frac{k}{M})}\right] - 0.5$
- For $x_i > \hat{F}^{-1}(1)$, $\hat{x}_i \leftarrow 1 - 0.5$.

Note that when the cdf has been estimated from the mini-batch, the first and third conditions should never occur.

$$y_i \leftarrow \gamma \hat{x}_i + \beta$$

Before moving on, we rewrite in somewhat simpler notation. For the following, let $g_k \equiv \hat{F}^{-1}(\frac{k}{M})$.

- For $x_i < g_0$, $\hat{x}_i \leftarrow 0 - 0.5$.

- For $g_k \leq x_i < g_{k+1}$,
  $\hat{x}_i \leftarrow \frac{1}{M+1}\left[k + \frac{x_i - g_k}{g_{k+1} - g_k}\right] - 0.5$

- For $x_i > g_M$, $\hat{x}_i \leftarrow 1 - 0.5$.

And now, mimicking the original batchnorm paper, we discuss some of the partial derivatives needed:

$$
\begin{aligned}
\frac{\partial l}{\partial \hat{x}_i} &= \frac{\partial l}{\partial y_i} \cdot \gamma \\
\frac{\partial \hat{x}_i}{\partial g_k} &= \frac{1}{M+1}\left[\frac{x_i - g_k}{(g_{k+1} - g_k)^2} - \frac{1}{g_{k+1} - g_k}\right], \quad g_k \leq x_i < g_{k+1} \\
\frac{\partial \hat{x}_i}{\partial g_{k+1}} &= \frac{-1}{M+1} \cdot \frac{x_i - g_k}{(g_{k+1} - g_k)^2}, \quad g_k \leq x_i < g_{k+1} \\
\frac{\partial \hat{x}_i}{\partial g_k} &= 0 \quad otherwise \\
\frac{\partial l}{\partial g_k} &= \sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial g_k} \\
\frac{\partial \hat{x}_i}{\partial x_i} &= \frac{1}{(M+1) \cdot (g_{k+1} - g_k)}, \quad g_k \leq x_i < g_{k+1} \\
\frac{\partial l}{\partial x_i} &= \frac{\partial l}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial x_i} + \sum_{k=0}^{M} \frac{\partial l}{\partial g_k} \frac{\partial g_k}{\partial x_i} \\
\frac{\partial l}{\partial \gamma} &= \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \cdot \hat{x}_i \\
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{m} \frac{\partial l}{\partial y_i}
\end{aligned}
$$

1.3 Calculating dxhatdg

For a given $\hat{x}_i$, $\frac{\partial \hat{x}_i}{\partial g_{gind}}$ is only non-zero for two values of $gind$: $gind_{left}$ and $gind_{right}$.

$gind_{left}$ is the percentile which is just less than (or equal to) x_i.

1.4 Calculating dgdx

Percentiles are estimated as a linear combination of two order statistics. However, in some cases, when the coefficient on the second order statistic is 0, they are effectively estimated from one order statistic.

Case 1: Estimating from 2 order statistics. For percentile k, the indices of the two order statistics that are used to estimate it are given by $zind$ and $zind + 1$, where

$$zind = \lfloor \frac{k \cdot (N+1)}{M} \rfloor.$$

Then, recall that

$$g_k = (1-d) \cdot z_{zind} + d \cdot z_{zind+1}.$$

Since each of these order statistics corresponds to exactly one original value of x, it follows that only two of the partial derivatives

$$\frac{\partial g_k}{\partial x_i}$$

will be non-zero.

In particular, it is for the x-indices that correspond to $zind$ and $zind + 1$. We shall refer to these as $xind_{left}$ and $xind_{right}$ since the values of x selected by these indices bracket the percentile on the left and the right. To retrieve these, we need a function which maps from a particular order statistic or "zind" back to the x index from which it came:

$$xind_{left} = xz[zind]$$

and

$$xind_{right} = xz[zind + 1].$$

Thus, starting with $\frac{\partial g_k}{\partial x_i}$ initialized to zero, we can set

$$\frac{\partial g_k}{\partial x_{xind_{left}}} = (1-d),$$

and

$$\frac{\partial g_k}{\partial x_{xind_{right}}} = d.$$

1.5 Calculating $\frac{\partial \hat{x}_i}{\partial x_i}$

Recall that

- For $g_k \le x_i < g_{k+1}$,
  $\hat{x}_i \leftarrow \frac{1}{M+1}\left[k + \frac{x_i - g_k}{g_{k+1} - g_k}\right] - 0.5.$

At the time that $\hat{x}_i$ is computed, we have the variables k, g_k, and g_{k+1} at hand, thus, we can easily calculate the derivative neded as:

$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{(M+1) \cdot (g_{k+1} - g_k)}, \quad g_k \le x_i < g_{k+1}. \tag{1}$$