AN ESTIMATOR FOR THE EFFECTIVE SEQUENCE WEIGHT EXPONENT

Question. Do sequences that get higher weight in training receive predictably better fits?

To measure how influenced by sequence weighting a model is, consider training (or fine-tuning) a model on training dataset $(X, Y, W) = \{(x_i, y_i, w_i)\}_{i\in\{1,2,\ldots,N\}}$ where w_i is a training sequence weight. We obtain trained model $\widehat{f}$ that we then generate in-sample predictions for: $\hat{y}_i := \widehat{f}(x_i)$ for $i \in \{1, 2, \ldots N\}$. We initially work in the large data setting (i.e. $N \to \infty$) and consider some set of weights W drawn from some bounded, nonnegative distribution $\mathcal{D}$.

Let $\mathcal{G} := \mathcal{G}_{\widehat{f}} : X \times \mathbf{R} \to \mathbf{R}$ measure the goodness of the fit of $\widehat{f}$ for feature x_i. We take $\mathcal{G}(x_i) =: g_i$ to be the *reduction of loss* (e.g. NLL) on sequence i by model $\widehat{f}$ versus a given sequence-weight-invariant baseline.

We define a population-level estimator for the optimal effective sequence weight exponent by choosing p^* to be the value of $p \geq 0$ that minimizes $d(E[G \mid W = w], w^p)$ over possible weights w for some distance-like quantity d .

Let (W, G, X, Y) describe a uniformly randomly selected training sequence after training our model (so W is the weight of the associated sequence), where G is the loss reduction associated to sequence (X, Y) and assume that $E[G] > 0$. To adjust for the lack of scale invariance of G and W, for a candidate exponent $p \geq 0$, define

$$Z_p := \frac{G}{\mathbf{E}[G]} - \frac{W^p}{\mathbf{E}[W^p]}.$$

By construction, $\mathbf{E}[Z_p] = 0$.

Rather than trying to define a loss over candidate $p \geq 0$ with the hope of obtaining an optimizing exponent p^* that exactly satisfies $\mathbf{E}[Z_{p^*} \mid W] \approx 0$, we take a smooth method-of-moments style relaxation of this restriction.

Given weight W, let $F_{\mathcal{D}}(W)$ be its cdf (i.e. percentile rank weight). Then for every $u \in [0, 1]$, let

$$M_p(u) := \mathbf{E}[Z_p \mathbf{1}_{\{F_{\mathcal{D}}(W)\leq u\}}].$$

If $\mathbf{E}[Z_p \mid W] = 0$, then $M_p(u) = 0$ for all u. Conversely if $M_p(u) = 0$ for all $u \in [0, 1]$ then almost everywhere $\mathbf{E}[Z_p \mid F_{\mathcal{D}}(W) = u] = 0$.

With this notation, we formally define p^* as the exponent that minimizes the following Cramér-von Mises-inspired cumulative discrepancy between G and W^p:

$$p^* := \operatorname{argmin}_{p\geq 0} \int_0^1 M_p(u)^2 du.$$

When G is nonnegative, this is an ordinary Cramér-von Mises distance between the two normalized distributions (but since G can be negative for individual sequences, this discrepancy is not necessarily a distance).

This resolves to a finite sample estimator by replacing the integral with an average over the training data. More precisely, let

$$\widetilde{g}_i := \frac{g_i}{\sum_{j\in\{1,2,\ldots,N\}} g_j}, \qquad b_i(p) := \frac{w_i^p}{\sum_{j\in\{1,2,\ldots,N\}} w_j^p}, \qquad d_i(p) := \widetilde{g}_i - b_i(p).$$

Let $\mathbf{d}(p) = (d_1(p), \ldots, d_N(p))^\top$, and let r_i be the empirical percentile rank of w_i. Define

$$K_{ij} := \min(r_i, r_j).$$

The finite-sample estimator is then

$$\widehat{p} = \operatorname*{argmin}_{p\geq 0} \mathbf{d}(p)^\top K \mathbf{d}(p).$$

Remark. On its face, an alternative estimator for p^* comes from the following nonlinear least squares optimization:

$$(\widehat{c}, \widehat{p}_{\mathrm{LS}}) = \operatorname*{argmin}_{c,\ p\geq 0} \sum_i (g_i - c \cdot w_i^p)^2,$$

If $\mathbf{E}[G \mid W = w] = c \cdot w^{p_0}$ holds exactly, both estimators in fact do recover $p^* = p_0$ asymptotically. Under misspecification they differ: least squares chooses the best pointwise approximation to the conditional mean, whereas our kernel estimator finds the best approximation to the allocation of explanatory contribution as measured by G across weight percentiles.

For any individual sequence, G is relatively noisy. Further, suppose that choosing exponent p' "incorrectly" ascribes $\widehat{f}$'s explanatory power to sequence i instead of sequence j. Estimating pointwise penalizes this error *independently* of the distance between w_i, w_j, whereas we might prefer to penalize the choice of p' more severely if the weights ascribed to i and j are very different.