Optimal sampling for imbalanced datasets

July 14, 2020

1 Statistical learning setting

Consider a random dataset $\mathcal{S}_N = \{(\mathbf{X}_i, Y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the target space. Note that we use bold notation denote a vector and capitalization to denote a random variable, a convention maintained throughout this paper. We make the assumption that $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_N, Y_N) \overset{\text{i.i.d.}}{\sim} p$, where the distribution p is unknown. Given a loss function $\psi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, we wish to find a hypothesis $f : \mathcal{X} \to \mathcal{Y}$ such that the expected loss

$$L(f) = \mathbb{E}_p \left[\psi(f(\mathbf{X}), Y) \right], \tag{1}$$

is minimized. However, as we only have access to the dataset $\mathcal{S}_N$ and not the underlying distribution p, we utilizes the empirical loss

$$\hat{L}(f) = \frac{1}{N} \sum_{i=1}^{N} \psi(f(\mathbf{X}_i), Y_i), \tag{2}$$

as an unbiased estimator of expected loss. We then solve the optimization problem

$$\min_{f \in \mathcal{F}} \hat{L}(f) + \lambda R(f), \tag{3}$$

where $\mathcal{F}$ is the hypothesis space, $R : \mathcal{F} \to \mathbb{R}$ is the regularization function and $\lambda \in \mathbb{R}_+$ the regularization parameter.

2 Variance reduction via importance sampling

We now discuss how to reduce the variance in the empirical loss $\hat{L}$ using an importance sampling approach, for a K-class classification problem with $\mathcal{Y} \in \{0, ..., K-1\}$. Given a valid probability distribution q with same support as p, the expected loss can be expressed as

$$L(f) = \mathbb{E}_q \left[\psi(f(\mathbf{X}), Y) \frac{p(\mathbf{X}, Y)}{q(\mathbf{X}, Y)} \right]. \tag{4}$$

The central insight of importance sampling is that a zero-variance estimator of L can be achieved under the optimal sampling density

$$q^*(\mathbf{x}, y; f) = \frac{\psi(f(\mathbf{x}), y) p(\mathbf{x}, y)}{L(f)}. \tag{5}$$

In most applications of importance sampling, the distribution p is known, allowing q^* to be approximated. However, in the statistical learning setting, estimating p is often a more challenging than the empirical loss minimization problem itself, especially for high-dimensional feature spaces. Nonetheless, for classification problems, it is straightforward to estimate the distribution of over the target space using the relative frequencies of each class in the dataset. Accordingly, if we then impose that

$$q_{\mathbf{X}|Y}(\mathbf{x}|y) = p_{\mathbf{X}|Y}(\mathbf{x}|y), \tag{6}$$

the importance sampling expression for the expected loss simplifies to

$$L(f) = \mathbb{E}_q \left[\psi(f(\mathbf{X}), Y) \frac{p_Y(Y)}{q_Y(Y)} \right], \tag{7}$$

which depends explicitly on probability distributions over the target space under p and q. The optimal sampling density is then given by

$$q_Y^*(y; f) = \underset{q \in \mathcal{D}_K}{\operatorname{argmin}} \ \operatorname{Var}_q \left(\psi(f(\mathbf{X}), Y) \frac{p_Y(Y)}{q_Y(Y)} \right), \tag{8}$$

where $\mathcal{D}_k$ is the set of valid probability distributions over the K classes.

We now present a simple approach to obtain a high-quality solution to (8). The objective can be decomposed as

$$\operatorname{Var}_q \left(\psi(f(\mathbf{X}), Y) \frac{p_Y(Y)}{q_Y(Y)} \right) = \mathbb{E}_q \left[\left(\psi(f(\mathbf{X}), Y) \frac{p_Y(Y)}{q_Y(Y)} \right)^2 \right] - L^2(f), \tag{9}$$

in which first term can be expressed as

$$\mathbb{E}_q \left[\left(\psi(f(\mathbf{X}), Y) \frac{p_Y(Y)}{q_Y(Y)} \right)^2 \right] = \mathbb{E}_p \left[\psi^2(f(\mathbf{X}), Y) \frac{p_Y(Y)}{q_Y(Y)} \right], \tag{10}$$

and the second term is independent of q, so it can be ignored. Therefore, given a realization of the dataset $\mathcal{S}_N$ drawn i.i.d. from p, the objective can be estimated as

$$\frac{1}{N} \sum_{i=1}^{N} \psi^2(f(\mathbf{x}_i), y_i) \frac{\hat{p}_Y(y_i)}{q_Y(y_i)}, \tag{11}$$

where:

$$\hat{p}_Y(k) \triangleq \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{y_i = k\}. \tag{12}$$

This results in the convex optimization problem

$$\begin{aligned}
&\min_{q_1, \ldots, q_k} \ \sum_{k=1}^{K} \frac{\hat{c}_k \hat{p}_k}{q_k}, && && (13)\\
&\sum_{k=1}^{K} q_k = 1, && && (14)\\
&q_k \le 1, && \forall k = 1, .., K, && (15)\\
&q_k \ge 0, && \forall k = 1, .., K, && (16)
\end{aligned}$$

where we have denoted $\hat{p}_k \triangleq \hat{p}_Y(k)$ and $q_k \triangleq q_Y(k)$, and defined

$$\hat{c}_k \triangleq \frac{1}{N} \sum_{i=1}^{N} \psi^2(f(\mathbf{x}_i), y_i) \mathbb{I}\{y_i = k\}. \tag{17}$$

For the case $K = 2$, this simplifies to

$$\min_{q \in [0,1]} \frac{\hat{c}_0(1 - \hat{p})}{1 - q} + \frac{\hat{c}_1 \hat{p}}{q}, \tag{18}$$

where $\hat{p}$ and q are the estimated nominal and sampling probabilities for the positive class. This has the optimal solution

$$q^* = \left(1 + \sqrt{\frac{\hat{c_0}(1 - \hat{p})}{\hat{c_1} \hat{p}}} \right)^{-1}. \tag{19}$$

When the loss function is the weighted (i.e. cost-sensitive) accuracy

$$\psi(f(\mathbf{x}), y) = \mathbb{I}\{y_i = 1 \vee f(\mathbf{x}) = 0\} \times \mathrm{FNC} + \mathbb{I}\{y_i = 0 \vee f(\mathbf{x}) = 1\} \times \mathrm{FPC}, \tag{20}$$

where FNC and FPC are respectively the false positive cost and false negative cost, the optimal sampling probability can be rewritten as

$$q^* = \left(1 + \left(\frac{1-\hat{p}}{\hat{p}}\right)\left(\frac{\mathrm{FPC}}{\mathrm{FNC}}\right)\sqrt{\frac{\mathrm{FPR}}{\mathrm{FNR}}}\right)^{-1} = 1 - \frac{1}{\gamma \cdot \mathrm{odds}(\hat{p}) + 1}, \tag{21}$$

where the error-adjusted cost ratio γ is defined as

$$\gamma \triangleq \frac{\mathrm{FNC}}{\mathrm{FPC}}\sqrt{\frac{\mathrm{FNR}}{\mathrm{FPR}}}, \tag{22}$$

and the odds function is given by

$$\mathrm{odds}(\hat{p}) = \frac{\hat{p}}{1-\hat{p}}. \tag{23}$$

We remark that this has the intuitive asymptotic behaviour that for a naturally balanced binary classification problem (i.e. $p = \frac{1}{2}$) with equal costs for false positives and negatives, $q^* = \frac{1}{2}$. Moreover, in the absence of an incumbent hypothesis f, a sensible initial guess for the optimal sampling probability, assuming equal error rates among the two classes, is given by:

$$q^* = 1 - \frac{1}{\frac{\mathrm{FNC}}{\mathrm{FPC}} \cdot \mathrm{odds}(\hat{p}) + 1}. \tag{24}$$

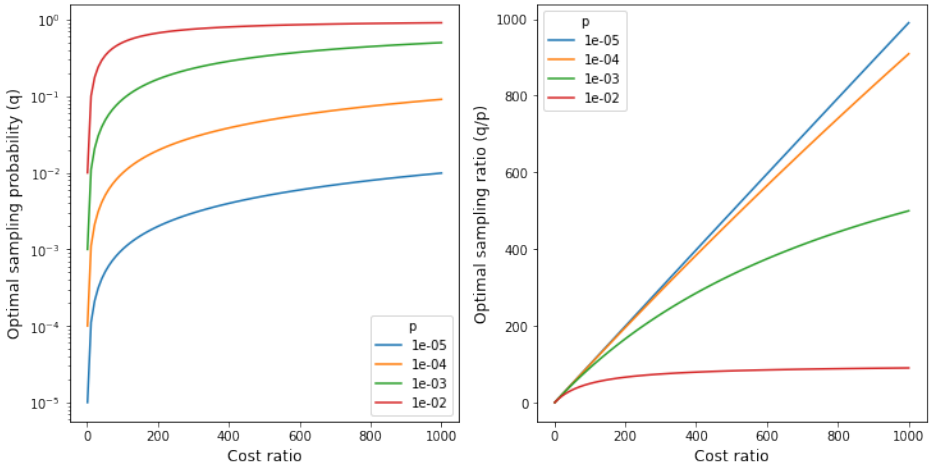

Figure 1: Optimal sampling behaviour for binary classification problem as function of the error-adjusted cost ratio γ, for various nominal probabilities p

3 Optimal undersampling

One limitation of the importance sampling approach is that it assumes minority class(es) can be oversampled, such that the dataset size N remains constant. In practice, however, dataset imbalance can often only be reduced by undersampling the majority class(es). We provide a modification of the above result for the binary classification case to address this. For a given sampling probability $q \geq \hat{p}$, the size of the undersampled dataset is

$$N' = \frac{\hat{p}N}{q}. \tag{25}$$

The variance in the empirical risk is then

$$\mathrm{Var}_q(\hat{L}) = \frac{q}{pN}\mathrm{Var}_q(L) \propto q\mathrm{Var}_q(L). \tag{26}$$

Discarding constant terms, this results the optimization problem

$$\min_{q\in[\hat{p},1]} \hat{c}_0(1-\hat{p})\frac{q}{1-q} - \hat{l}^2 q, \tag{27}$$

where $\hat{l}$ is the realized empirical loss. This has the optimal solution

$$q^* = \max\left\{1 - \frac{\sqrt{\hat{c}_0(1-\hat{p})}}{\hat{l}}, \hat{p}\right\} \tag{28}$$

which can be expressed for the weighted accuracy loss function as

$$q^* = 1 - \frac{1}{\gamma \cdot \mathrm{odds}(\hat{p}) \cdot \sqrt{\mathrm{FNR}} + \sqrt{\mathrm{FPR}}}. \tag{29}$$

Note that

$$\frac{1}{\gamma \cdot \mathrm{odds}(\hat{p}) \cdot \sqrt{\mathrm{FNR}} + \sqrt{\mathrm{FPR}}} \geq \frac{1}{\gamma \cdot \mathrm{odds}(\hat{p}) + 1}, \tag{30}$$

meaning the minority (positive) class is sampled with a lower probability than in the importance sampling case to reduce data wastage.

4 Algorithm

We operationalize the key result developed in Section 3, (), via a coordinate descent method outlined in Algorithm. 1. This approach overcomes the complication that the optimal sampling probability is dependedent on the hypothesis function by iteratively updating both the model parameters and sampling proability until convergence. Cross-validation is employed to estimate the updated sampling probability at each iteration to reduce variance, and updates are clipped to prevent the oscillations and promote stable convergence behaviour.

Algorithm 1: `OptimalUndersampling`

```
input  : Dataset $\mathcal{S}_N$
         Hypothesis space $\mathcal{F}$ parameterized by $\theta \in \mathbb{R}^L$
         Loss function $\psi$
         Number of cross-validation folds $K$
         Max iterations $T$
         Max update $\Delta$
         Termination toleration $\delta$
$\hat{p} \leftarrow \frac{1}{N} \sum_{(\mathbf{x},y) \in \mathcal{S}_N} \mathbb{I}\{y = 1\}$
$q^{(0)} \leftarrow \frac{1}{2}$
for $t$=1,...,$T$ do
    for $k$=1,...,$K$ do
        $\mathcal{S}^{(\text{train})} \leftarrow \texttt{Resample}\left(\mathcal{S}_N \setminus \text{Fold } k, q^{(t-1)}\right)$
        $\mathcal{S}^{(\text{val})} \leftarrow \text{Fold } k$
        $\theta^{(t)} \leftarrow \underset{\theta \in \mathbb{R}^L}{\operatorname{argmin}} \frac{1}{|\mathcal{S}^{(\text{train})}|} \sum_{(\mathbf{x},y) \in \mathcal{S}^{(\text{train})}} \psi(f(\mathbf{x};\theta), y) \frac{\hat{p}y + (1-\hat{p})(1-y)}{q^{(t-1)}y + (1-q^{(t-1)})(1-y)}$
        $\hat{c}_0 \leftarrow \frac{1}{|\mathcal{S}^{(\text{val})}|} \sum_{(\mathbf{x},y) \in \mathcal{S}^{(\text{val})}} \psi^2(f(\mathbf{x};\theta^{(t)}), y) \mathbb{I}\{y = 0\}$
        $\hat{l} \leftarrow \frac{1}{|\mathcal{S}^{(\text{val})}|} \sum_{(\mathbf{x},y) \in \mathcal{S}^{(\text{val})}} \psi(f(\mathbf{x};\theta^{(t)}), y)$
        $q^{(t,k)} \leftarrow 1 - \frac{\sqrt{\hat{c}_0(1-\hat{p})}}{\hat{l}}$
    end
    $q^{(t)} \leftarrow \text{clip}\left(\frac{1}{K} \sum_{k=1}^{K} q^{(t,k)}, q^{(t-1)} - \Delta, q^{(t-1)} + \Delta\right)$
    if $\left|q^{(t)} - q^{(t-1)}\right| < \delta$ then
        break
    end
end
output: Optimal sampling probability $q^{(t)}$
```