Learning Feature Discretizations

November 18, 2022

To deal with continuous features within our online learning framework, we had to "binarize" the features. In particular, for each feature we consider different thresholds for binarization, and in each training time step we select the threshold that maximizes the gain based on our criterion. More formally, for each random variable (feature) X_i we assume that there exist K different binary latent random variables $\{Z_{i1}, Z_{i2}, \ldots, Z_{iK}\}$, and each of these latent random variables corresponds to a threshold for binarizing X_i. Assume that given a label Y_j, the random variable Z_{ik} is distributed by a Bernoulli distribution with parameter $\theta_{ij}^{(k)} = \mathbb{P}[X_i, Z_{ik} = 1 \mid Y_j]$. As before, we assume that we have some prior information about $\theta_{ij}^{(k)}$ in the form of a prior (Beta) distribution.

In each time step t, we start by sampling from the posterior distribution of parameters $\theta_{ij}^{(k)}$ for all i, j, k. Then, we seek the feature that maximizes the feature query score. In the case of continuous features, we need to additionally find the best binarization threshold for each feature; this can be done by computing the gains achieved with each threshold and selecting the one that maximizes the gain. But this exhaustive way of selecting the thresholds is computationally expensive, and we do not exploit the information available from our data. To improve this process, we can view the threshold selection procedure (for each feature) as an adversarial bandit problem with arms and rewards being the set of thresholds and marginal gains respectively. Here the randomness comes from the features in the data point.

Let $\Pi_{ti} : [K] \to \mathbb{R}^+$ ($\sum_k \Pi_{ti}(k) = 1$) be the probability distribution according to which we select the binarization threshold for X_i at time step t. Then, threshold selection and updating Π_{ti} can be done with the following procedure which is adapted from the Exp3 algorithm:

Algorithm 1

```
Require: $\eta$
  $S_{ik}^{(0)} \leftarrow 0$ for all $k$
  for each time step $t$ do
    if feature $i$ is queried then
      Calculate the threshold sampling distribution:
          $\Pi_{ti}(k) = \frac{\exp(\eta S_{ik}^{(t-1)})}{\sum_{k'} \exp(\eta S_{ik'}^{(t-1)})}$
      Sample threshold $B_{ti} \sim \Pi_{ti}$
      Observe marginal gain $\Delta_{ti}$
      Calculate $S_{ik}^{(t)}$:
          $S_{ik}^{(t)} = S_{ik}^{(t-1)} + \frac{\mathbb{I}\{B_{ti}=k\}\Delta_{ti}}{\Pi_{ti}(k)}$
    end if
  end for
```

The above procedure can be used with or without feature selection.