AVE BIAS EVALUATION

BRIAN DAVIS

1. Introduction

Let F be a feature space (e.g. $\mathbb{R}^n$), let $X \subset F$ be a collection of binary labeled data, and let $\sigma : F \times F \to [0,1]$ be a similarity distance measure (e.g., Jaccard/Tanimoto distances). Let V denote the validation set and T denote the training set, with V_N (resp. T_N) from the negative class and V_P (resp. T_P) from the positive class, so that $X = (V_P \sqcup T_P) \bigsqcup (V_N \sqcup T_N)$.

2. AVE bias

For v in V, define the function $I_d(v,T)$ to be equal to one if $\min_{t\in T}\{\sigma(v,t)\} < d$ and zero otherwise.

Wallach et al. [1] describe a bias metric for evaluating training/validation splits in binary classification problems. Define the function $H_{(V,T)}$ by

$$H_{(V,T)} = \frac{1}{n} \cdot \frac{1}{|V|} \sum_{v\in V} \left(\sum_{i=1}^{n} I_{i/n}(v,T) \right).$$

Then the Asymmetric Validation Embedding (AVE) bias is defined[1] to be the quantity

$$\begin{aligned} B(V_P,V_N,T_P,T_N) &= H_{(V_P,T_P)} - H_{(V_P,T_N)} + H_{(V_N,T_N)} - H_{(V_N,T_P)} \\ &= \frac{1}{|V_P|} \cdot \frac{1}{n} \sum_{v\in V_P} \left[\left(\sum_{i=1}^{n} I_{i/n}(v,T_P) \right) - \left(\sum_{i=1}^{n} I_{i/n}(v,T_N) \right) \right] \\ &+ \frac{1}{|V_N|} \cdot \frac{1}{n} \sum_{v\in V_N} \left[\left(\sum_{i=1}^{n} I_{i/n}(v,T_N) \right) - \left(\sum_{i=1}^{n} I_{i/n}(v,T_P) \right) \right]. \end{aligned}$$

Notice that for t in T we have $\sigma(v,t) < i/n$ if and only if $\lfloor n\sigma(v,t) \rfloor \leq i-1$, where $\lfloor z \rfloor$ is the greatest integer less than or equal to z. Thus $I_{i/n}(v,T)$ is equal to one if and only if $\lfloor n \min_{t\in T}\{\sigma(v,t)\} \rfloor \leq i-1$.

Note that $I_{i/n}(v,T)$ equal to one implies that $I_{j/n}(v,t)$ equals one for $i \leq j \leq n$, so that

$$\sum_{i=1}^{n} I_{i/n}(v,T) \;=\; n - \lfloor n \min_{t\in T}\{\sigma(v,t)\} \rfloor.$$

For simplicity, we write

$$\Gamma(v,T) = \frac{\left\lfloor n \cdot \min_{t\in T_N}\{\sigma(v,t)\} \right\rfloor - \left\lfloor n \cdot \min_{t\in T_P}\{\sigma(v,t)\} \right\rfloor}{n},$$

Date: January 2019.

[1]The original definition used $1/(n+1)$ in place of the $1/n$ term, which we find more convenient.

and so

$$B(V_P, V_N, T_P, T_N) = \text{mean}_{v \in V_P}\{\Gamma(v, T)\} - \text{mean}_{v \in V_N}\{\Gamma(v, T)\}. \tag{1}$$

Ideally, a bias measure would be robust against slight perturbations of the data. Unfortunately, since the floor function is not continuous, perturbations of v do not translate into perturbations of the AVE bias. Instead, the bias changes in discrete "jumps", with size bounded by $1/n$. A better definition of the AVE bias may be to take the limit as $n \to \infty$, which is well defined for finite data sets. Thus, practically, it is sufficient to use a very large value of n because the complexity ($\mathcal{O}(|V| \cdot |T|)$ as computed above) of computing the AVE bias is insensitive to n. A value of $n = 100$ was given in [1] in the definition of AVE bias, and a value of 50 was used in the accompanying code of that paper.

Since $0 \leq z - \lfloor z \rfloor < 1$ for all z, we have that

$$0 \leq n \cdot \min_{t \in T}\{\sigma(v, t)\} - \left\lfloor n \cdot \min_{t \in T}\{\sigma(v, t)\}\right\rfloor < 1,$$

and in the limit as n goes to infinity,

$$\Gamma(v, T) = \min_{t \in T_N}\{\sigma(v, t)\} - \min_{t \in T_P}\{\sigma(v, t)\},$$

so that

$$\begin{aligned} B(V_P, V_N, T_P, T_N) = \text{mean}_{v \in V_P}\Big\{ &\min_{t \in T_N}\{\sigma(v, t)\} - \min_{t \in T_P}\{\sigma(v, t)\}\Big\} \\ + \ &\text{mean}_{v \in V_N}\{\min_{t \in T_P}\{\sigma(v, t)\} - \min_{t \in T_N}\{\sigma(v, t)\}\}. \end{aligned} \tag{2}$$

Looking at Equation 2, we see that for a given distribution of training data, the AVE bias approximates a measure of the frequency with which validation points of a given class are nearer (in the σ sense) to training points of the same class than to training points of the opposite class, and to what extent.

References

[1] Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of chemical information and modeling*, 58(5):916–932, 2018.

E-mail address: Brian.Davis@uky.edu