Derivation of Bias-Adjusted Metrics (used in the official results notebook)

Let $T(x)$ be the biased selection process for labelling data at DOHMH. We treat it as a black-box and model it atomically.

Let U be the set of all Yelp Reviews that have been processed by the system.

Let $B \subset U$ s.t. $B = \{(x, y)|T(x) = 1\}$. B is the biased Yelp review set, labeled by DOHMH epidemiologists after selection by the prototype classifier.

Let $B^c \subseteq U \setminus B = \{(x, y)|T(x) = 0\}$. B^c is the complement of the biased set: all reviews which were never seen by DOHMH epidemiologists because they were filtered out by the prototype classifier.

Error Rate

Let $\overline{B} \subseteq B$ be the labeled points from B which are seen in the *training* data.

Likewise, let $\overline{B}^c \subset B^c$ be the sample of labeled points from B^c which are seen in the *training* data.

We can model the error rate of some classifier f as:

$$p(f(x) \neq y) = p(f(x) \neq y|T(x) = 1)p(T(x) = 1)+p(f(x) \neq y|T(x) = 0)p(T(x) = 0)$$

and use plugin estimates:

$$\hat{p}(T(x) = 1) = \frac{|B|}{|U|} \quad \text{and} \quad \hat{p}(T(x) = 0) = 1 - \frac{|B|}{|U|}$$

$$\hat{p}(f(x) \neq y|T(x) = 1) = \frac{1}{|\overline{B}|} \sum_{(x,y)\in\overline{B}} I[f(x) \neq y]$$

$$\hat{p}(f(x) \neq y|T(x) = 0) = \frac{1}{|\overline{B}^c|} \sum_{(x,y)\in\overline{B}^c} I[f(x) \neq y]$$

therefore

$$\hat{p}(f(x) \neq y) = \frac{1}{|\overline{B}|}\frac{|B|}{|U|}\sum_{(x,y)\in\overline{B}} I[f(x) \neq y] + \frac{1}{|\overline{B}^c|}\Big(1 - \frac{|B|}{|U|}\Big)\sum_{(x,y)\in\overline{B}^c} I[f(x) \neq y]$$

Since in practice we will average the errors over the entire training set, we multiply and divide the quantity by $|\bar{B}| + |\bar{B^c}|$, yielding:

$$ErrorRate = \frac{1}{|\overline{B}|+|\overline{B}^c|}\Big[w_{\overline{B}}\sum_{(x,y)\in\overline{B}} I[f(x) \neq y] + w_{\overline{B}^c}\sum_{(x,y)\in\overline{B}^c} I[f(x) \neq y]\Big]$$

where

$$w_{\overline{B}} = \frac{|\overline{B}| + |\overline{B}^c|}{|\overline{B}|}\frac{|B|}{|U|} \quad \text{and} \quad w_{\overline{B}^c} = \frac{|\overline{B}| + |\overline{B}^c|}{|\overline{B}^c|}\Big(1 - \frac{|B|}{|U|}\Big)$$

Testing

We will calculate the weights as was done in the above Error Rate calculation, however this time we must recalculate the weights using the observed test data proportions.

So, let $\underline{B} \subseteq B$ be the labeled points from B which are seen in the *test* data.

Likewise, let $\underline{B}^c \subset B^c$ be the sample of labeled points from B^c which are seen in the *test* data.

Then we have

$$w_{\underline{B}} = \frac{|\underline{B}| + |\underline{B}^c|}{|\underline{B}|}\frac{|B|}{|U|} \quad \text{and} \quad w_{\underline{B}^c} = \frac{|\underline{B}| + |\underline{B}^c|}{|\underline{B}^c|}\Big(1 - \frac{|B|}{|U|}\Big)$$

and precision and recall can be calculated as follows:

In the following equations, let $(x_i, y_i) \in \underline{B} \cup \underline{B}^c$

Precision

$$Precision = \frac{1}{\sum\limits_{f(x_i)=1} w_i}\Big(\sum_{f(x_i)=1} w_i I[y_i = 1]\Big)$$

Recall

$$Recall = \frac{1}{\sum\limits_{y_i=1} w_i}\Big(\sum_{y_i=1} w_i I[f(x_i)=1]\Big)$$

F1

Using the above plugin estimates to calculate the importance weighted precision and recall, we can caluclate the bias-adjusted F1-score using:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

AUPR

Finally, we can obtain a series of ordered bias-adjusted Precision-Recall points $E = \{(p,r)_i | r_i \geq r_{i'}, i' < i\}$ by varying the classification threshold $t \in (1, 0]$ and then using trapezoidal integration to approximate the Area Under the Recall vs. Precision curve (AUPR).

$$AUPR = \sum_{(p_i,r_i)\in E, i<|E|} .5 * (r_{i+1} - r_i) * (p_i + p_{i+1})$$

Bootstrap

For the final evaluation, we would like confidence intervals about the F1 and AUPR. We find these by using the percentile bootstrap.

We calculate bootstrap statistics for the IW-Precision, IW-Recall, and IW-AUPR as follows:

First we calculate the statistic $\bar{x}$ (for each of IW-Precision, IW-Recall, and IW-AUPR).

Then we resample the test dataset with replacement B times and obtain the bootstrap statisic estimates for each set.

Call these $x_1, ..., x_B$.

Then we can compute confidence intervals around $\bar{x}$ the usual way by finding the $\alpha = .025$ boundary quantiles $\delta_\alpha, \delta_{1-\alpha}$, such that

$$P(\bar{x}^* - \delta_{1-\alpha} \leq \bar{x} \leq \bar{x}^* - \delta_\alpha) = .95$$