Supplementary Materials for "Reactive Learning: Active Learning with Relabeling"

1 Theorem 1

Proof. We first show that the theorem is true when $\mathcal{X}_L$ only contains singly-labeled examples. $\texttt{US}^{\alpha}_{\mathcal{X}}$ will always pick an unlabeled example x_u over a singly-labeled example x_l, if α is set such that $(1-\alpha)M_A(x_l)+\alpha M_L(x_l) < (1-\alpha)M_A(x_u)+\alpha M_L(x_u)$ for all x_l, x_u pairs. This condition holds true when $\alpha > \frac{M_A(x_l)-M_A(x_u)}{M_A(x_l)-M_A(x_u)+M_L(x_u)-M_L(x_l)}$ for all x_l, x_u pairs. We set $\alpha' = \sup_{x_l\in\mathcal{X}_L, x_u\in\mathcal{X}_L} \frac{0.69}{0.69+(M_L(x_u)-M_L(x_l))}$. Note that since x_l is singly-labeled and will have lower label entropy compared to x_u, which is unlabeled, $M_L(x_u) > M_L(x_l)$. Therefore, $\alpha' < 1.0$. Also, since M_A is an entropy of a binary random variable, $|M_A(x_l) - M_A(x_u)| < 0.69$. Combining all these facts, the condition holds true when $\alpha > \alpha' > \frac{0.69}{0.69+(M_L(x_u)-M_L(x_l))}$ for all x_l, x_u and $\alpha < \frac{0.69}{0.69-(M_L(x_u)-M_L(x_l))} > 1.0$ for all x_l, x_u. Since all unlabeled examples have the same label uncertainty and because $\texttt{US}^{\alpha}_{\mathcal{X}}$ always picks an unlabeled example, the example it picks will be determined based on the classifier's uncertainty, just as in $\texttt{US}_{\mathcal{X}_U}$. Now, since both $\texttt{US}^{\alpha}_{\mathcal{X}}$ and $\texttt{US}_{\mathcal{X}_U}$ start with $\mathcal{X}_L = \emptyset$, by induction, $\mathcal{X}_L$ will only ever contain singly-labeled examples, and so these two strategies are equivalent. $\square$

2 Theorem 2

Let $P_{\mathcal{A}}(h^*(x_i) = y)$ denote the probability currently output by learning algorithm, $\mathcal{A}$, that $h^*(x_i) = y$. For ease of notation and clarity, we denote with shorthand $p_0(x_i) = P_{\mathcal{A}}(h^*(x_i) = 0)$ and $p_1(x_i) = P_{\mathcal{A}}(h^*(x_i) = 1)$. Because we are considering a setting with no noise, the total expected impact of a point x_i is $\sum_{y\in\mathcal{Y}} p_y(x_i)\psi_y(x_i)$.

Lemma 1. *If*

1. $(\psi_1(x_i) - \psi_0(x_i)) \geq \frac{\psi_0(x_j)-\psi_0(x_i)+(\psi_1(x_j)-\psi_0(x_j))p_1(x_j)}{p_1(x_i)}$, *or*

2. $(\psi_0(x_i) - \psi_1(x_i)) \geq \frac{\psi_1(x_j)-\psi_1(x_i)+(\psi_0(x_j)-\psi_1(x_j))p_0(x_j)}{p_0(x_i)}$, *or*

3. $(\psi_0(x_i) - \psi_1(x_i)) \geq \frac{\psi_0(x_j)-\psi_1(x_i)+(\psi_1(x_j)-\psi_0(x_j))p_1(x_j)}{p_0(x_i)}$, *or*

4. $(\psi_1(x_i) - \psi_0(x_i)) \geq \frac{\psi_1(x_j)-\psi_0(x_i)+(\psi_0(x_j)-\psi_1(x_j))p_0(x_j)}{p_1(x_i)}$,

then, the total expected impact of x_i is at least as large as that of x_j: $\sum_{y\in\mathcal{Y}} p_y(x_i)\psi_y(x_i) \geq \sum_{y\in\mathcal{Y}} p_y(x_j)\psi_y(x_j)$.

Proof. For condition (1), we have that $\sum_{y\in\mathcal{Y}} p_y(x_i)\psi_y(x_i)$

$$
\begin{aligned}
&= p_0(x_i)\psi_0(x_i) + p_1(x_i)\psi_1(x_i) \\
&= p_1(x_i)(\psi_1(x_i) - \psi_0(x_i)) + \psi_0(x_i) \\
&\geq p_1(x_i)\frac{\psi_0(x_j) - \psi_0(x_i) + (\psi_1(x_j) - \psi_0(x_j))p_1(x_j)}{p_1(x_i)} + \psi_0(x_i) \\
&= \psi_0(x_j) - \psi_0(x_i) + (\psi_1(x_j) - \psi_0(x_j))p_1(x_j) + \psi_0(x_i) \\
&= \psi_0(x_j) + (\psi_1(x_j) - \psi_0(x_j))p_1(x_j) \\
&= \sum_{y\in\mathcal{Y}} p_y(x_j)\psi_y(x_j).
\end{aligned}
$$

Proofs of conditions (2-4) proceed similarly. □

2.1 Proof of Theorem 2

Proof. Let x_i be the point chosen by uncertainty sampling. We prove the theorem by showing that $\mathcal{P}$ satisfies the conditions of Lemma 1 for x_i and all candidate points x_j. We prove the case when $x_i > t$ and $x_j > t$ (then $x_j > x_i$, because otherwise x_i would not have been picked by uncertainty sampling). The 3 other cases proceed in exactly the same manner, because of symmetry. Let us also assume that $x_i < x_<$, because if not, the theorem holds trivially, because all points will have 0 impact. Let t be the currently learned threshold, $x_< = \max\{x \in \mathcal{X}_L : x < t\}$ denote the current greatest labeled example less than the threshold, and $x_> = \min\{x \in \mathcal{X}_L : x > t\}$ denote the current smallest labeled example greater than the threshold. Now we define $d_{*_1,*_2}$ to be the proportion of points in $\mathcal{X}$ between points $*_1$ and $*_2$. Precisely,

$$d_{*_1,*_2} = P_{x\in\mathcal{D}}(x \in \{x : *_1 < x < *_2\}).$$

For example, $d_{x_<,t}$ is the proportion of points between $x_<$ and t. We also know that $d_{x_j,t} \geq d_{x_i,t}$ because $x_j > x_i$. Now we show that condition 1 of Lemma 1 is satisfied, that $(\psi_0(x_i) - \psi_1(x_i)) \geq \frac{\psi_0(x_j)-\psi_0(x_i)+(\psi_1(x_j)-\psi_0(x_j))p_1(x_j)}{p_1(x_i)}$ for all $x_j > x_i$

We have that for any x_j, $\psi_0(x_j) = d_{t,x_j} + \frac{d_{x_j,x_>}}{2}$ and $\psi_1(x_j) = d_{x_<,x_j} - (\frac{d_{x_<,x_j}}{2} + d_{t,x_j})$. Therefore, $\psi_1(x_j) - \psi_0(x_j)$

$$
\begin{aligned}
&= d_{x_<,x_j} - (\frac{d_{x_<,x_j}}{2} + d_{t,x_j})) - (d_{t,x_j} + \frac{d_{x_j,x_>}}{2}) \\
&= d_{x_<,t} - \frac{d_{x_<,x_j}}{2} - \frac{d_{x_j,x_>}}{2} - d_{t,x_j} \\
&= d_{x_<,t} - d_{d_<,t} - d_{t,x_j} \\
&= -d_{t,x_j}.
\end{aligned}
$$

Next, we have that $\frac{\psi_0(x_j)-\psi_0(x_i)+(\psi_1(x_j)-\psi_0(x_j))p_1(x_j)}{p_1(x_i)}$

$$
\begin{aligned}
&= \frac{d_{t,x_j} + \frac{d_{x_j,x_>}}{2} - (d_{t,x_i} + \frac{d_{x_i,x_>}}{2}) - d_{t,x_j}p_1(x_j)}{p_1(x_i)} \\
&= \frac{d_{x_i,x_j} - 0.5(d_{x_i,x_j}) - d_{t,x_j}p_1(x_j)}{p_1(x_i)} \\
&= \frac{0.5d_{x_i,x_j} - d_{t,x_j}p_1(x_j)}{p_1(x_i)}.
\end{aligned}
$$

And then,

$$
\begin{aligned}
\frac{0.5d_{x_i,x_j} - d_{t,x_j}p_1(x_j)}{p_1(x_i)} &\leq -d_{t,x_i} = (\psi_1(x_i) - \psi_0(x_j)) \\
&\Updownarrow \\
0.5d_{x_i,x_j} - d_{t,x_j}p_1(x_j) &\leq -d_{t,x_i}p_1(x_i) \\
&\Updownarrow \\
0.5d_{x_i,x_j} &\leq d_{t,x_j}p_1(x_j) - d_{t,x_i}p_1(x_i) \\
&\Updownarrow \\
0.5d_{x_i,x_j} &\leq d_{t,x_j}[p_1(x_i) + \beta] - d_{t,x_i}p_1(x_i), \quad \beta = p_1(x_j) - p_1(x_i) \\
&\Updownarrow \\
0.5d_{x_i,x_j} &\leq p_1(x_i)d_{x_i,x_j} + \beta d_{t,x_j}, \quad \beta = p_1(x_j) - p_1(x_i)
\end{aligned}
$$

$\beta > 0$ because $x_j > x_i$, and $p_1(x_i) > 0.5$ because $x_i > t$, and therefore $0.5d_{x_i,x_j} \leq p_1(x_i)d_{x_i,x_j} + \beta d_{t,x_j}$, and the theorem is proved.

$\square$