(atbegshi) Package atbegshi Warning: Ignoring void shipout box

6. Appendix

6.1. Proof of Claim 1

Proof The expected conditional risk can be solved optimally by a dynamic program, where a DP recursion is,

$$J_K(x^K, S^K) = \min_{f^K} \mathbf{E}_y \left[R_k(y, x^k, f^k, S^k) \right] \tag{23}$$

$$J_k(x^k, S^k) = \min_{f^k} \mathbf{E}_y \left[R_k(y, x^k, f^k, S^k) \right] + \ \mathbf{E}_{x^{k+1} \dots x^K} \left[J_{k+1}(x^{k+1}, S^{k+1}) \Big| \, x^k \right] \tag{24}$$

Consider kth stage minimization, f^k can take three possible values $+1, -1, r$ and $J_k(x^k, S^k)$ can be recast as an conditional expected risk minimization,

$$J_k(x^k, S^k) = S^k \min_{f^k} \left\{ \underbrace{w_n \mathrm{P}_y \left[y = -1 \mid x \right]}_{f^k = +1}, \underbrace{w_p \mathrm{P}_y \left[y = +1 \mid x \right]}_{f^k = -1}, \underbrace{\delta^k + \mathbf{E}_{x^{k+1} \dots x^K} \left[J_{k+1}(x^{k+1}, S^{k+1} = 1 \right]}_{f^k = r} \right\} \tag{25}$$

Now, define $\tilde{\delta}(x^k) = \delta^k + \mathbf{E}_{x^{k+1} \dots x^K} \left[J_{k+1}(x^{k+1}, S^{k+1} = 1) \right]$ and solve the conditional risk above for $S^k = 1$,

$$f^k(x^k) = \begin{cases} -1, & \text{if } \mathrm{P}(y = 1 | x^k) < \frac{\tilde{\delta}(x^k)}{w_p} \\ r, & \text{if } \frac{\tilde{\delta}(x^k)}{w_p} \leq \mathrm{P}(y = 1 | x^k) \leq 1 - \frac{\tilde{\delta}(x^k)}{w_n} \\ +1, & \text{if } \mathrm{P}(y = 1 | x^k) > 1 - \frac{\tilde{\delta}(x^k)}{w_p} \end{cases} \tag{26}$$

which is exactly our claim. ■

6.2. Proof of Claim 2

Proof The conditional expected risk for a given x, $\tilde{\delta}(x)$ and error penalties w_n,w_p is,

$$\min \left\{ \underbrace{w_n \mathrm{P}_y \left[y = -1 \mid x \right]}_{f = +1}, \underbrace{w_p \mathrm{P}_y \left[y = +1 \mid x \right]}_{f = -1}, \underbrace{\tilde{\delta}(x)}_{f = r} \right\} \tag{27}$$

The optimal bayesian classifier that minimizes this risk is,

$$f(x) = \begin{cases} -1, & \text{if } \mathrm{P}(y = 1 | x) < \frac{\tilde{\delta}(x)}{w_p} \\ r, & \text{if } \frac{\tilde{\delta}(x)}{w_p} \leq \mathrm{P}(y = 1 | x) \leq 1 - \frac{\tilde{\delta}(x)}{w_n} \\ +1, & \text{if } \mathrm{P}(y = 1 | x) > 1 - \frac{\tilde{\delta}(x)}{w_p} \end{cases} \tag{28}$$

We need to show that f can be decomposed as a pair of binary classifiers $f_n, f_p : \mathcal{X} \to \{+1, -1\}$. Consider the following decomposition,

$$f(x) = \begin{cases} f_p(x), & f_p(x) = f_n(x) \\ r, & f_p(x) \neq f_n(x) \end{cases} \tag{29}$$

The conditional expected risk with this decomposition,

$$\min \left\{ \underbrace{w_n \mathrm{P}_y\left[y = -1 \mid x\right]}_{f_p(x)=+1, f_n(x)=+1}, \underbrace{w_p \mathrm{P}_y\left[y = +1 \mid x\right]}_{f_p(x)=-1, f_n(x)=-1}, \quad \underbrace{\tilde{\delta}(x)}_{f_p(x)\neq f_n(x)} \right\} \tag{30}$$

Note that the expected risk is symmetric and f_n and f_p can be interchanged. However, consider the equations for f_p and f_n that follow from minimizing the risk. Here, we used the fact that $\mathrm{P}(y = -1|x) = 1 - \mathrm{P}(y = 1|x)$.

$$f_p(x) = \begin{cases} +1, & \mathrm{P}(y = 1|x) > \frac{\tilde{\delta}(x)}{w_p} \\ -1, & \mathrm{P}(y = 1|x) \leq \frac{\tilde{\delta}(x)}{w_p} \end{cases} \tag{31}$$

$$f_n(x) = \begin{cases} +1, & \mathrm{P}(y = 1|x) > 1 - \frac{\tilde{\delta}(x)}{w_n} \\ -1, & \mathrm{P}(y = 1|x) \leq 1 - \frac{\tilde{\delta}(x)}{w_n} \end{cases} \tag{32}$$

Note that we chose our convention such that f_p is positively biased classifier and f_n is negatively biased classifier.

And, by inspection, 29 is true, therefore our decomposition is the optimal bayesian classifier.

Also, note another interesting observation, f_p and f_n are solutions to the following biased classification problems,

$$f_p = \arg\min \left\{ \underbrace{\left(1 - \frac{\tilde{\delta}(x)}{w_p}\right) \mathrm{P}_y\left[y = -1 \mid x\right]}_{f=+1}, \underbrace{\left(\frac{\tilde{\delta}(x)}{w_p}\right) \mathrm{P}_y\left[y = +1 \mid x\right]}_{f=-1} \right\} \tag{33}$$

$$f_n = \arg\min \left\{ \underbrace{\left(\frac{\tilde{\delta}(x)}{w_n}\right) \mathrm{P}_y\left[y = -1 \mid x\right]}_{f=+1}, \underbrace{\left(1 - \frac{\tilde{\delta}(x)}{w_n}\right) \mathrm{P}_y\left[y = +1 \mid x\right]}_{f=-1} \right\} \tag{34}$$

$$\tag{35}$$

Here, we used a standard Bayesian solution to a conditional expected risk for binary classification with weights k and $1 - k$,

$$f^* = \arg\min \left\{ \underbrace{(1 - k)\, \mathrm{P}_y\left[y = -1 \mid x\right]}_{f=+1}, \underbrace{(k)\, \mathrm{P}_y\left[y = +1 \mid x\right]}_{f=-1} \right\} \tag{36}$$

$$\tag{37}$$

$$f^*(x) = \begin{cases} +1, & \mathrm{P}(y=1|x) > k \\ -1, & \mathrm{P}(y=1|x) \leq 1-k \end{cases} \tag{38}$$

■

6.3. Proof of Theorem 1

Proof Since the risk is a smooth function of $\mathbf{q}_n, \mathbf{q}_p, \mathbf{q}^2$, our algorithm solves the following by coordinate descent minimization over $\mathbf{q}_n, \mathbf{q}_p, \mathbf{q}^2$:

$$\min_{\mathbf{q}_n, \mathbf{q}_p, \mathbf{q}^2} \hat{R}(f_n, f_p, f^2) \tag{39}$$

$$s.t. f_p = \sum_{h_j \in \mathcal{H}^1} q_j^p h_j(x_i), \;\; f_n = \sum_{h_j \in \mathcal{H}^1} q_j^n h_j(x_i) \tag{40}$$

$$f^2 = \sum_{h_j \in \mathcal{H}^2} q_j^2 h_j(x_i) \tag{41}$$

therefore we are guaranteed to converge to a local minimum. ■

6.4. Theorem 2 (Generalization Error Bound)

Our approach employs margin maximizing algorithm.(Masnadi-Shirazi and Vasconcelos (2009)) So it is appropriate to prove an error margin generalization bound for a two stage system:

Theorem 2 *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{+1, -1\}$, and let $\mathcal{S}$ be a sample of m examples chosen independently at random according to $\mathcal{D}$, and a rejected subsample of size m_r, $\mathcal{S}_r = \{x \in \mathcal{S} | f_p(x) \neq f_n(x)\}$ Assume that the base-classifier spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are finite, and let $\delta > 0$. Then with probability at least $1-\delta$ over the random choice of the training set S, all boosted classifiers f_n, f_p, f_2 satisfy the following bound for all $\theta_1 > 0$ and $\theta_2 > 0$:*

$$\begin{aligned} \mathrm{P}_{\mathcal{D}}[yf_n(x) \leq 0, yf_p(x) \leq 0] + \mathrm{P}_{\mathcal{D}}[yf_2(x) \leq 0, f_n(x) \neq f_p(x)] \leq \\ \mathrm{P}_{\mathcal{S}}[yf_n(x) \leq \theta_1, yf_p(x) \leq \theta_1] + \mathrm{P}_{\mathcal{S}_r}[yf_2(x) \leq \theta_2] \\ + \mathcal{O}\left(\frac{1}{\sqrt{m}}\left(\frac{\log m \log |\mathcal{H}_1|}{\theta_1} + \log\frac{1}{\delta}\right)^{\frac{1}{2}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{m_r}}\left(\frac{\log m_r \log |\mathcal{H}_2|}{\theta_2} + \log\frac{1}{\delta}\right)^{\frac{1}{2}}\right) \end{aligned} \tag{42}$$

Proof This will closely follow the proof of Theorem 1 in Bartlett et al. (1998). We have to bound two terms: $\mathrm{P}_{\mathcal{D}}[yf_n(x) \leq \theta_1, yf_p(x) \leq \theta_1]$ and $\mathrm{P}_{\mathcal{D}}[yf_2(x) \leq \theta_2, yf_n(x) \neq yf_p(x)]$

First Term Let us bound the first term. Define $\mathcal{C}_N$ to be the set of unweighted averages over N elements from $\mathcal{H}_1$,

$$\mathcal{C}_N = \{f : x \rightarrow \frac{1}{N}\sum_{i=1}^{N} h_i(x) \mid h_i \in \mathcal{H}_1\} \tag{43}$$

Any weighed classifier $f = \sum_h q_h h(x)$ can be approximated by drawing an element from $\mathcal{C}_N$ by choosing $h_1 ... h_N$ with prob. q_h.

We can express our first term as a sum of probabilities of disjoint events.

$$\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0\right] = \tag{44}$$

$$\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0, y g_p(x) \leq \frac{\theta_1}{2}, y g_n(x) \leq \frac{\theta_1}{2}\right] \tag{45}$$

$$+\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0, y g_p(x) \leq \frac{\theta_1}{2}, y g_n(x) > \frac{\theta_1}{2}\right] \tag{46}$$

$$+\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0, y g_p(x) > \frac{\theta_1}{2}, y g_n(x) \leq \frac{\theta_1}{2}\right] \tag{47}$$

$$+\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0, y g_p(x) > \frac{\theta_1}{2}, y g_n(x) > \frac{\theta_1}{2}\right] \tag{48}$$

Further, we can write,

$$\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0\right] \leq \mathrm{P}_{\mathcal{D}}\left[y g_p(x) \leq \frac{\theta_1}{2}, y g_n(x) \leq \frac{\theta_1}{2}\right] \tag{49}$$

$$+\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0, y g_p(x) > \frac{\theta_1}{2}, y g_n(x) > \frac{\theta_1}{2}\right] \tag{50}$$

The inequality holds for any g_p, g_n. We take the expected value of the right hand side wrt to the distribution $\mathcal{C}$

$$\mathrm{P}_{\mathcal{D}}\left[y f_p(x) \leq 0, y f_n(x) \leq 0\right] \leq \tag{51}$$

$$\mathbf{E}_{\mathcal{C}}\left[\mathrm{P}_{\mathcal{D}}\left[y g_p(x) \leq \frac{\theta_1}{2}, y g_n(x) \leq \frac{\theta_1}{2}\right]\right] \tag{52}$$

$$+\mathbf{E}_{\mathcal{D}}\left[\mathrm{P}_{\mathcal{C}_p,\mathcal{C}_n}\left[y g_p(x) > \frac{\theta_1}{2}, y g_n(x) > \frac{\theta_1}{2} \mid y f_p(x) \leq 0, y f_n(x) \leq 0\right]\right] \tag{53}$$

The last term inside the expectation is the probability that an average of N bernoulli random variables is larger than its expectation, we use a concentration result from Equation (4) in Theorem 1 of Bartlett et al. (1998).

$$\mathrm{P}_{\mathcal{C}_p,\mathcal{C}_n}\left[y g_p(x) > \frac{\theta_1}{2}, y g_n(x) > \frac{\theta_1}{2} \mid y f_p(x) \leq 0, y f_n(x) \leq 0\right] \leq exp\left(\frac{-N\theta_1^2}{8}\right) \tag{54}$$

To bound the first we use the result from Equation (5) in Theorem 1 of Bartlett et al. (1998). if we set $\epsilon_N = \sqrt{(1/2m)\log((N+1)|\mathcal{H}_1|^{2N})/\delta_N}$, with probability at least $1 - \delta_N$,

$$\mathrm{P}_{\mathcal{D},\mathcal{C}}\left[y g_p(x) \leq \frac{\theta_1}{2}, y g_n(x) \leq \frac{\theta_1}{2}\right] \leq \mathrm{P}_{S,\mathcal{C}}\left[y g_p(x) \leq \frac{\theta_1}{2}, y g_n(x) \leq \frac{\theta_1}{2}\right] + \epsilon_N \tag{55}$$

for any choice of θ and every distribution $\mathcal{C}$. Here, $\mathrm{P}_S\left[\right]$ is probability taken with respect to a randomly drawn sample of size m from $\mathcal{D}$.

By the same argument as in inequality 50,

$$\mathrm{P}_{S,\mathcal{C}_p}\left[yg_p(x) \le \frac{\theta_1}{2}, yg_n(x) \le \frac{\theta_1}{2}\right] \le \tag{56}$$

$$\mathrm{P}_S\left[yf_p(x) \le \theta_1, yf_n(x) \le \theta_1\right] + \mathbf{E}_S\left[\mathrm{P}_{\mathcal{C}_p}\left[yg_p(x) \le \frac{\theta_1}{2} \mid yf_p(x) > \theta\right]\right] \tag{57}$$

The expressions inside the expectation can be bounded using the same Chernoff bound result from 54,

$$\mathrm{P}_{\mathcal{C}}\left[yg_p(x) \le \frac{\theta_1}{2}, yg_n(x) \le \frac{\theta_1}{2} \mid yf_p(x) > \theta_1, yf_p(x) > \theta_1\right] \le exp\left(\frac{-N\theta_1^2}{8}\right) \tag{58}$$

By setting $\delta_N = \delta/(N(N+1))$, and combining the terms,

$$\mathrm{P}_{\mathcal{D}}\left[yf_p(x) \le 0, yf_n(x) \le 0\right] \le \tag{59}$$

$$\mathrm{P}_S\left[yf_p(x) \le \theta_1, yf_n(x) \le \theta_1\right] + 2exp\left(\frac{-N\theta_1^2}{8}\right) + 2\sqrt{\frac{1}{2m}\log\left(\frac{N(N+1)^2|\mathcal{H}_1|^{2N}}{\delta}\right)} \tag{60}$$

By setting, $N = (4/\theta_1^2)\log(m/\log|\mathcal{H}_1|^2)$,

$$\mathrm{P}_{\mathcal{D}}\left[yf_p(x) \le 0, yf_n(x) \le 0\right] \le \mathrm{P}_S\left[yf_p(x) \le \theta_1, yf_n(x) \le \theta_1\right] + \mathcal{O}\left(\frac{1}{\sqrt{m}}\left(\frac{\log m \log|\mathcal{H}|^2}{\theta} + \log\frac{1}{\delta}\right)^{\frac{1}{2}}\right) \tag{61}$$

Second Term Here we will bound the second term, $\mathrm{P}_{\mathcal{D}}[yf_2(x) \le \theta_2, yf_n(x) \ne yf_p(x)]$ Define a new distribution:

$$D_r = \begin{cases} cD(x,y), & f_p(x) \ne f_n(x) \\ 0, & f_p(x) = f_n(x) \end{cases} \tag{62}$$

Rewrite:

$$\mathrm{P}_{\mathcal{D}}[yf_2(x) \le \theta_2, yf_n(x) \ne yf_p(x)] \le \mathrm{P}_{\mathcal{D}}[yf_2(x) \le \theta_2 \mid yf_n(x) \ne yf_p(x)] \tag{63}$$

$$= \mathrm{P}_{\mathcal{D}_r}[yf_2(x) \le \theta_2] \tag{64}$$

Note that $\mathcal{S}_r$ is an iid sample from $\mathcal{D}_r$. Using Theorem 1 in Bartlett et al. (1998),

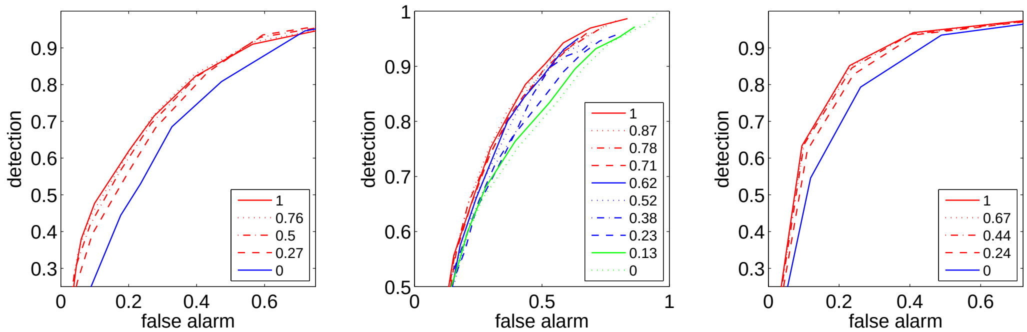
$$\mathrm{P}_{\mathcal{D}_r}[yf_2(x) \le 0] \le \mathrm{P}_{\mathcal{S}_r}[yf_2(x) \le \theta_2] + \mathcal{O}\left(\frac{1}{\sqrt{m}}\left(\frac{\log m \log|\mathcal{H}_2|}{\theta_2} + \log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)$$

Collecting the two terms produces the desired result. ■

Discussion: The error is a sum of two terms: first stage error of data that is not rejected and second stage error on rejected fraction. It states that if we are given a first and second stage boosted classifiers than we can bound the generalization error by the empirical margin error over the training set and a term that is inversely proportional to the margin and the number of training examples at that stage. An interesting observation is that m_r depends on the reject classifier at first stage. So if very few examples make it to the second stage then we do not have strong generalization.

6.5. Additional Experiments

In medical diagnosis and threat detection, the penalty of false positives and false negatives (w_n, w_p) is not equal. The experiment in Fig. 10 demonstrates our global algorithms in the biased scenario. For each reject cost δ, we compute an ROC curve. We also compute a corresponding average reject rate for each value of delta. This reject rate is averaged over the values (w_n, w_p). So the highest reject rate corresponds to the best performance but also to the highest acquisition cost incurred by the system.



(a) pima (b) polyps (c) threat

Figure 10: Two Stage ROC using the global surrogate method. Each ROC curve corresponds to a different value of reject cost δ. The legend displays average reject rate for δ's. Note, the red ROC corresponds to the centralized system (100% reject rate). For both experiments, very good performance can be achieved by requesting only 50% of instances to be measured at the second stage.