Appendix: Learn to Expect the Unexpected: Probably Approximately Correct Domain Generalization

A Proving bounds on correlation coefficients

This section includes the proof of Lemma 6.

Proof of Lemma 6. Let ρ and $\hat{\rho}$ be the correlation coefficient and empirical correlation of R, S on a sample of size m. Let $q_{ij} = \Pr[R = i \wedge S = j]$ and $\hat{q}_{ij}$ be the corresponding realized empirical fractions over the m samples.

By Chernoff bounds, for any i, j, the probability that $|q_{ij} - \hat{q}_{ij}| > \tau$ is at most $2e^{-2m\tau^2} \leq \delta/4$ for $\tau = \epsilon^2 v/64$. Hence, with probability $\geq 1 - \delta$, $|q_{ij} - \hat{q}_{ij}| \leq \tau$ for and for all i, j. We now argue that if this happens, then $|\rho - \hat{\rho}| \leq \epsilon$.

As shorthand, let $a = q_{00}, b = q_{01}, c = q_{10}, d = q_{11}$ and $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ be the analogous empirical quantities. It may be helpful for the reader to draw a 2x2 table of possible values of R, S and associated probabilities.

Case 1: $c + d \leq \tau$. In this case we use $|\rho - \hat{\rho}| \leq |\rho| + |\hat{\rho}|$ and argue that both $|\rho|, |\hat{\rho}| \leq 2\sqrt{\tau/v} \leq \epsilon/2$. To see this, the definition of correlation coefficient applied to binary random variables means that correlation can be written as

$$\rho = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{1}$$

and similarly for $\hat{\rho}$. Since all quantities are non-negative, we can remove terms to get

$$-\sqrt{\frac{c}{a}} \leq \frac{-bc}{\sqrt{(b)(c)(a)(b)}} \leq \rho \leq \frac{ad}{\sqrt{(a)(d)(a)(b)}} = \sqrt{\frac{d}{b}},$$

and similarly for $\hat{\rho}$. In turn this implies that $|\rho| \leq \max\{\sqrt{c/a}, \sqrt{d/b}\}$. Since $c + d \leq \tau$, we have that $c, d \leq \tau$ and since $\mathbb{E}[S] \in [v, 1 - v]$ we have that $a + c, b + d \geq v$, in turn implying $a, b \geq v - \tau$. Hence,

$$|\rho| \leq \max\left\{\sqrt{\frac{c}{a}}, \sqrt{\frac{d}{b}}\right\} \leq \sqrt{\frac{\tau}{v - \tau}}.$$

Similarly, for $\hat{\rho}$, we have

$$|\hat{\rho}| \leq \max\left\{\sqrt{\frac{\hat{c}}{\hat{a}}}, \sqrt{\frac{\hat{d}}{\hat{b}}}\right\} \leq \max\left\{\sqrt{\frac{c+\tau}{a-\tau}}, \sqrt{\frac{d+\tau}{b-\tau}}\right\} \leq \sqrt{\frac{\tau+\tau}{v-\tau-\tau}} \leq \sqrt{\frac{2\tau}{v-2\tau}}.$$

This upper bound is greater than the one we have for $|\rho|$. Hence,

$$|\rho - \hat{\rho}| \leq |\rho| + |\hat{\rho}| \leq 2\sqrt{\frac{2\tau}{v-2\tau}} \leq 2\sqrt{\frac{2\tau}{v/2}} = 4\sqrt{\frac{\tau}{v}} \leq \epsilon.$$

In the above we have used the fact that $2\tau \leq v/2$.

Case 2: $c+d \in [\tau, 1/2]$. We use the fact that, given that $|\hat{a}-a|, |\hat{b}-b| \le \tau$,

$$\frac{\hat{a}}{\hat{a}+\hat{b}} \le \frac{a+\tau}{(a+\tau)+(b-\tau)} = \frac{a+\tau}{a+b}, \tag{2}$$

because $\alpha/(\alpha+\beta)$ is increasing in α and decreasing in β. From (1), one can see that

$$\rho = \sqrt{\frac{a}{a+b}\cdot\frac{d}{c+d}\cdot\frac{a}{a+c}\cdot\frac{d}{b+d}} - \sqrt{\frac{b}{a+b}\cdot\frac{c}{c+d}\cdot\frac{c}{a+c}\cdot\frac{b}{b+d}}. \tag{3}$$

The bounds on $\hat{a}/(\hat{a}+\hat{b})$ imply that

$$\begin{aligned}\sqrt{\frac{\hat{a}}{\hat{a}+\hat{b}}\cdot\frac{\hat{d}}{\hat{c}+\hat{d}}\cdot\frac{\hat{a}}{\hat{a}+\hat{c}}\cdot\frac{\hat{d}}{\hat{b}+\hat{d}}} &\le \sqrt{\frac{a+\tau}{a+b}\cdot\frac{d+\tau}{c+d}\cdot\frac{a+\tau}{a+c}\cdot\frac{d+\tau}{b+d}} \\ &= \frac{(a+\tau)(d+\tau)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \\ &\le \frac{ad+2\tau}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \end{aligned} \tag{4}$$

In the last step above, we used the fact that $\tau(a+d) \le \tau$ since $a+d \le 1$. Similarly to (2), we have

$$\frac{\hat{a}}{\hat{a}+\hat{b}} \ge \frac{\max\{0, a-\tau\}}{a+b}.$$

Combining with a similar lower bound to (4) gives

$$\left| \frac{\hat{a}\hat{d}}{\sqrt{(\hat{a}+\hat{b})(\hat{c}+\hat{d})(\hat{a}+\hat{c})(\hat{b}+\hat{d})}} - \frac{ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \right| \le \frac{2\tau}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Applying the same argument replacing ad with bc and substituting into (3) gives

$$|\hat{\rho}-\rho| \le \frac{4\tau}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

By assumption $c+d \le a+b$ hence $a+b \ge 1/2$ and $(a+b)(c+d) \ge \tau/2$, and since $\mathbb{E}[S] = b+d \in [v, 1-v]$, we have $(a+c)(b+d) \ge v/2$. Combining with the above gives $|\hat{\rho}-\rho| \le 8\sqrt{\tau/v} = \epsilon$.

Case 3: $c+d \ge 1/2$. Replacing R by $1-R$ negates ρ and also negates $\hat{\rho}$. This transformation swaps a with c and b with d but preserves $|\rho-\hat{\rho}|$. Hence, we can use the prior two cases which cover $c+d \le 1/2$. □

B Additional Details: Decision Tree Multi-Domain Model

Proof of Theorem 7. For high-probability bounds, it suffices to guarantee *expected* error rate at most $\epsilon\delta$, for $m, d \ge q(\frac{sn}{\epsilon\delta})$, for some polynomial q, by Markov's inequality.

First, it is not difficult to see that the algorithm will never have any false positives, i.e., it will never predict positively when the true label is negative. To see this, note that each positive prediction must arise because of at least one c_i. As mentioned above, Kearns 1994 show that the largest consistent classifier with any set of (noiseless) positive data is conservative in that it never has any false positives. Hence the above algorithm will never have any false positives.

We bound the expected rate of false negatives (which is equal to the expected error rate) by summing over leaves and using linearity of expectation. False negatives in positive leaf ℓ can arise in two ways: (a) leaf ℓ was simply never chosen as a domain, and (b) leaf ℓ was chosen $z^i = \ell$ for some $i \le d$, but there is a term

$x[j] = k$ for some $j \leq n, k \in \{0, 1\}$ which occurs in c_i but not in z_ℓ in which case any positive example that satisfies $x[j] = k$ will be a false negative. Moreover, these are the only types of false negatives. Hence, the expected rate of false negatives coming from leaf ℓ with probability p_ℓ due to (a) is $p_\ell(1 - p_\ell)^d$, the fraction of examples from leaf ℓ times the probability that domain z_ℓ was never chosen. The expected rate of false negatives due to (b) is at most $p_\ell 2n/(m + 1)$, again the probability of leaf ℓ times $2n/(m + 1)$. To see why, note that there are at most $2n$ terms ($x[j] = k$) not in the true conjunction z_ℓ and, for each such term, the expected error contribution can be upper bounded by imagining picking $m + 1$ examples at random, m for training and 1 for test. The probability that among $m + 1$ positive examples, that only example which would satisfy that term would be the one chosen for test is $1/(m + 1)$. Hence the expected rate of false negatives and hence also the expected error rate is at most

$$\sum_\ell p_\ell(1 - p_\ell)^d + p_\ell \frac{2n}{m + 1} < \frac{s}{d} + \frac{2n}{m}. \tag{5}$$

The inequality above holds for the left term because $r(1 - r)^d \leq 1/d$ for $r \in [0, 1]$ and for the right term because the $\sum p_\ell = 1$ and for the left term by concavity of $\sum p_\ell(1 - p_\ell)^d$ on the probability simplex. Note that the above error rate is bounded by $\epsilon\delta$ if we have $d \geq 4s/(\epsilon\delta)$ and $m \geq 4n/(\epsilon\delta)$, which completes our proof. □