Practical Ensemble Classification Error Bounds for Different Operating Points

PROBLEM SETUP

1. Supervised binary classification $y \in \{-1, 1\},\ \mathrm{x} \in \mathcal{X}$
2. Ensemble classification: $\phi(\mathrm{x}) = \frac{1}{m}\sum_i^m \hat{y}_i(\mathrm{x}) \in [-1, 1]$
3. $\hat{y}(\mathrm{x}) = \begin{cases} -1 \ \text{ if } \phi(\mathrm{x}) \leq 0 \\ 1 \ \text{ if } \phi(\mathrm{x}) > 0 \end{cases}$

Classification with a Reject Option

$$\hat{y}(\mathrm{x}) = \begin{cases} -1 \ \text{ if } \phi(\mathrm{x}) \leq t \\ \text{reject, if } \phi(\mathrm{x}) \in (-t, t) \\ 1 \ \text{ if } \phi(\mathrm{x}) \geq t \end{cases}$$

- rejection, provides a guard band around the decision region

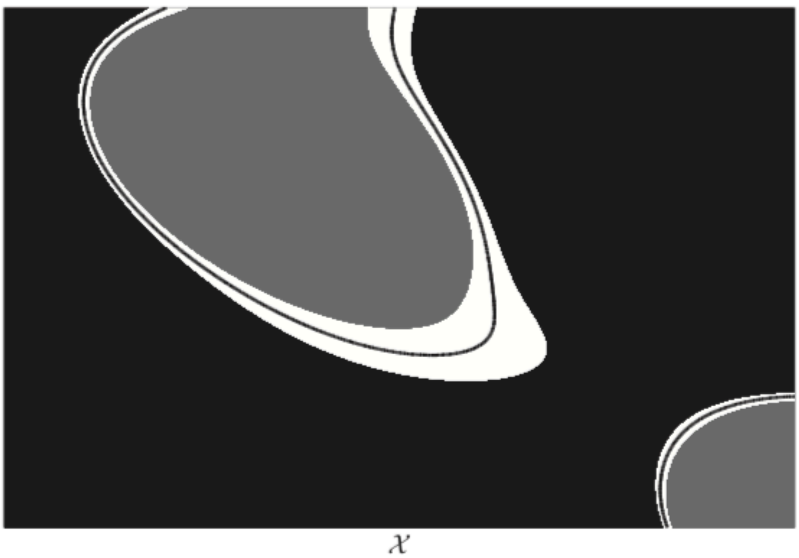

Fig. 1. Illustration of decision regions in feature space. The region $\hat{y} = +1$ is black, the region $\hat{y} = -1$ is gray, the region $\hat{y} =$ reject is white, and the boundary $\phi = 0$ is the black line.

- margin z measures the extent to which the average number of votes for the right class exceeds the average vote for any other class.

$$z = mr(\mathrm{x}, y) = av_k I(\hat{y}_k(\mathrm{x}) = y) - \max_{j \neq Y} av_k I(\hat{y}_k(\mathrm{x}) = j)$$

  - Error Probability: $P_E(t) = P_{X,Y}(mg(x, y) < -t)$

- if y=1 $mg(x,y) = av_1 - av_{-1} = \phi(x) < -t$
- if y=-1 $mg(x,y) = av_{-1} - av_1 = -\phi(x)$
  $\phi(x) > t \Rightarrow z < -t$

  - Rejection Prabability $P_R(t) = \Pr[z \in (-t,t)]$

- measure of performance： reject option risk

$$L_c(t) = \underbrace{P_E(t)}_{\text{Cost of misclassification}} + \underbrace{cP_R(t)}_{\text{cost of rejection is c}}$$

- Classication Error:
  - False Alarm : False Positive (FP) $\hat{y} = +1 \wedge y = -1$

$$P_F(t) = \Pr[\phi > t | y = -1] = \int_t^1 f(\phi | y = -1) d\phi$$

  - Missed dectection : False Negative (FN) $\hat{y} = -1 \wedge y = +1$

$$P_M(t) = \Pr[\phi \leq t | y = +1] = \int_{-1}^t f(\phi | y = +1) d\phi$$

- The dectection probability : the percentage of alarms can be detected

$$P_D(t) = \Pr[\phi > t | y = +1] = \int_t^1 f(\phi | y = +1) d\phi$$

BOUNDS BASED ON STRENGTH AND CORRELATION

- correlation:

$$\bar{p} = \frac{2}{m(m-1)} \sum_{i \neq j} \mathbb{E}[\hat{y}_i(\mathrm{x}) \hat{y}_j(\mathrm{x})]$$

- strength : $s = \mathbb{E}[z]$

From [8] the random forest paper, it is true that

$$\mathrm{var}(z) = \mathbb{E}[(z-s)^2] \leq \bar{p}(1-s^2)$$

Suppse s>0, better than randomly predict, than the following bound on generalization error is derived in [8] using the Chebyshev inequality,

$$\Pr(y \neq \hat{y}(x)) \leq \frac{\bar{p}(1-s^2)}{s^2}$$

Chebyshev Inequality

Let X be a random variable for which Var(X) exists, then for every t>0,

$$\Pr(|X - \mathbb{E}[X]| > t) \leq \frac{\mathrm{Var(X)}}{t^2}$$

Bound for Reject Option Risk

Based on the Cantelli (one-sided Chebyshev) inequality

- $P_E(t) \leq \frac{1}{1+\frac{(s+t)^2}{\bar{p}(1-s^2)}}, s > -t$
- $\Pr[z < t] \leq \frac{1}{1+\frac{(s-t)^2}{\bar{p}(1-s^2)}}, s > t$

$$\begin{aligned} L_c(t) = P_E(t) + cP_R(t) &= \Pr[z \leq -t] + c\Pr[-t \leq z < t] \\ &= (1-c)\Pr[z \leq -t] + c\Pr[z < t] \\ &= (1-c)P_E(t) + c\Pr[z < t] \\ &\leq \frac{1-c}{1+\frac{(s+t)^2}{\bar{p}(1-s^2)}} + \frac{c}{1+\frac{(s-t)^2}{\bar{p}(1-s^2)}},\ s > t \end{aligned}$$

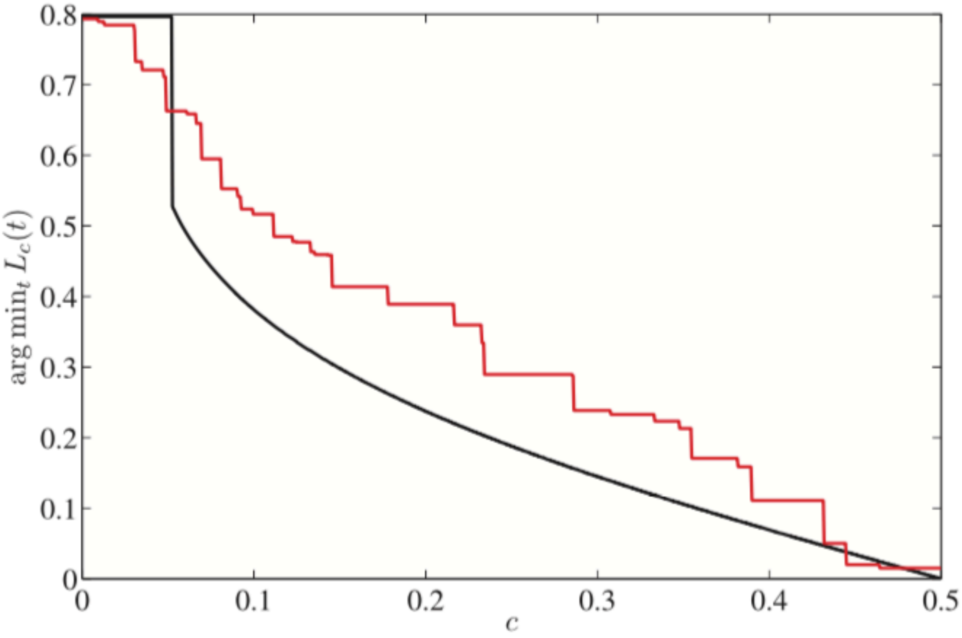

Fig. 11. Rejection threshold that minimizes risk as a function of rejection cost, empirically (red line) and on the analytical bound (black line) for the spambase data set.

Bound for Receiver Operating Characteristic

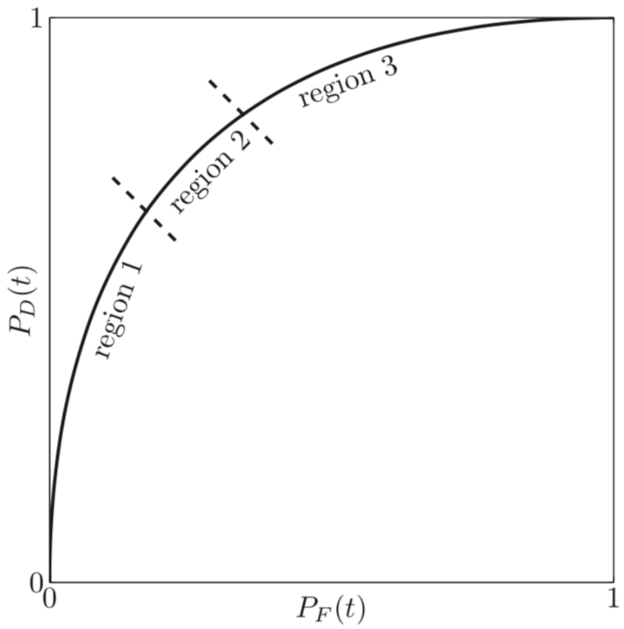

Fig. 5. Illustration of ROC split into Region 1: $t \in [s_+, 1]$, Region 2: $t \in [-s_-, s_+]$, and Region 3: $t \in [-1, -s_-]$.

Conditional correlation

- $\bar{p}_+ = \frac{2}{m(m-1)} \sum_{i \neq j} \mathbb{E}[\hat{y}_i(\mathrm{x})\hat{y}_j(\mathrm{x})|y=+1]$
- $\bar{p}_- = \frac{2}{m(m-1)} \sum_{i \neq j} \mathbb{E}[\hat{y}_i(\mathrm{x})\hat{y}_j(\mathrm{x})|y=-1]$

Conditional strength

- $s_+ = \mathbb{E}[\phi|y=+1]$
- $s_- = -\mathbb{E}[\phi|y=-1]$
- $s = s_+\mathrm{Pr}[y=1] + s_-\mathrm{Pr}[y=-1]$

The bound for dectection probability : the percentage of alarms can be detected,

$$P_D(t) = \mathrm{Pr}[\phi > t|y=+1] \leq \frac{1}{1+\frac{(s_+-t)^2}{\bar{p}_+(1-s_+^2)}}, s_+ < t$$

$$P_D(t) \geq \frac{1}{1+\frac{\bar{p}_+(1-s_+^2)}{(s_+-t)^2}}, s_+ > t$$

False alarm bound:

$$P_F(t) \leq \frac{1}{1+\frac{(s_--t)^2}{\bar{p}_-(1-s_-^2)}}, -s_- < t$$

$$P_F(t) \geq \frac{1}{1+\frac{\bar{p}_-(1-s_-^2)}{(s_--t)^2}}, -s_- > t$$

The implicit bound

If $t \in [-s_-, s_+]$,, we have $P_D(t) \geq \frac{1}{1+\frac{\bar{p}_+(1-s_+^2)}{(s_+-t)^2}}$, and $P_F(t) \leq \frac{1}{1+\frac{(s_--t)^2}{\bar{p}_-(1-s_-^2)}}$

- if $t = -s_-$, $P_F \leq 1$
- if $t = s_+$, $P_D > 0$

For small false alarm probability, $\bar{p}_-$ should be small and s_- should be large. For large detection probability, $\bar{p}_+$ should be small and s_+ should be large.

$$P_D \geq \begin{cases} 0 \text{ if } P_F \leq \frac{\eta_F}{\eta_F+1} \\ \frac{1}{1+\eta_m\left(1-\sqrt{\eta_F(P_F^{-1}-1)}\right)^{-2}} \text{ if } P_F > \frac{\eta_F}{\eta_F+1} \end{cases}$$

$$\eta_m = \frac{\bar{p}_+(1-s_+^2)}{(s_+s_-)^2}$$

$$\eta_F = \frac{\bar{p}_-(1-s_-^2)}{(s_+s_-)^2}$$

To push the ROC up in the low missed detection regime, we would like η_m to be as close to zero as possible.