Instance → (continuous risk) score s

Cost ratio c

decision/classification threshold t

$$cost(t) = FP(t) + cFN(t)$$

↳ total cost of all decisions when t is applied

Minimize the expected cost by taking the derivative w.r.t. t and setting it to zero:

$f(s)$ → density of cases w/ a given score

$E[Y|s]$ → rate of cancer cases (positives) among the patients for a given score

$$\frac{dE[cost(t)]}{dt} = \frac{dE[FP(t)]}{dt} + c\frac{dE[FN(t)]}{dt}$$

$$0 = -f(t^*)E[1-Y|s=t^*] + cf(t^*)E[Y|s=t^*]$$

↳ positive

↑ threshold: ↓ FPs, ↑ FNs

$f(t^*)$ → # cases at the threshold

$E[Y|s=t^*]$ → fraction of cases that were negative

$$t^*_{impl} = E[Y|s=t^*] = \frac{1}{1+c}$$

↳ implied threshold

$\frac{1}{1+1} = \frac{1}{2} = 0.5$

score at which cases have a $1/(1+c)$ probability of being positive.

$$FNR = \frac{FN}{\underbrace{FN+TP}_{P}}$$

$\frac{3}{3+4} = \frac{3}{7} \simeq 0.43$ male

$\frac{2}{2+7} = \frac{2}{9} \simeq 0.22$ female

group a → publisher
group b → consumer

$$cost_{ab} = FP_{ab}(t_{ab}) + c_{ab}FN_{ab}(t_{ab})$$

$$c = \frac{1-E[Y|s=t]}{E[Y|s=t]}$$ → cost ratio → implied threshold

$E[Y|s=t, female] = 0.25$

$c = \frac{1-0.25}{0.25} = 3$

$E[Y|s=t, male] = 0.16$

$c = (1-0.16)/0.16 \simeq 5$

FN

Decision threshold t

Score s

Label $\hat{Y}$

$\hat{Y}_i = \mathbb{1}\{s_i \geq t\}$ ← indicator function

$t_{implied} = E[Y|s=t]$ → prob. of the outcome corresponding to the decision threshold

If (classifier) calibrated, $s = E[Y|s=s]$, then the decision and implied thresholds are identical

$E[Y|s=t] \approx E[Y|t_{lower} \leq s \leq t_{upper}]$, $t_{low} \leq t < t_{upper}$

↓ → implied threshold estimation

$Y_i = \beta_0 + \beta_1 (s_i - t) + \varepsilon_i$ weighted linear regression

$$w_i = \begin{cases} \left[1 - \left(\frac{|s_i - t|}{d}\right)^3\right]^3 & s_i \in [t-d, t+d] \\ 0 & \text{otherwise} \end{cases}$$

↳ half-width, symmetric region around the threshold

$$FPR = \frac{FP}{FP+TN}$$

Fairness measurement dataset:

(Y_{ij}, Y_i^*) → expert-provided ground-truth for that post

↙ label provided by labeler j to post i

Signal Detection Theory (SDT):

Parameters: ϕ prevalence

d' separation → distance between the Gaussian's means.

$$FPR(d', t) = 1 - \Phi(t)$$

$$FNR(d', t) = \Phi(t - d')$$

model parameters ↙ ↓ cumulative dist. function for the standard normal

using the observed error rates: $\phi \rightarrow$ prevalence

$$t = \Phi^{-1}(1 - FPR)$$

$$d' = t - \Phi^{-1}(FNR)$$

↳ inverse cumulative dist. function for the standard normal

(Prevalence and Bayes' rule)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$$E[Y|s=t] = P(Y=1|s=t)$$

$$= \frac{P(Y=1)\,P(s=t|Y=1)}{P(Y=1)\,P(s=t|Y=1) + P(Y=0)\,P(s=t|Y=0)}$$

$\rightarrow$ prevalence

$$= \frac{\phi N(t-d')}{\phi N(t-d') + (1-\phi) N(t)}$$

Equal variance

$N(\cdot) \rightarrow$ density function of the standard normal dist.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

implied threshold

$$c = \frac{1 - E[Y|s=t]}{E[Y|s=t]} \quad (\div \phi N(t-d'))$$

$$= \frac{1}{1 + \frac{1-\phi}{\phi} \exp(-td' + d'^2/2)}$$

$$c = \frac{1-\phi}{\phi} \exp\left(-td' + d'^2/2\right)$$

$$-\frac{1}{2}\left(\frac{x-t}{\sigma}\right)^2 - \left(-\frac{1}{2}\left(\frac{x-(t-d')}{\sigma}\right)^2\right)$$

$$-\frac{1}{2}(x-t)^2 + \frac{1}{2}(x-(t-d'))^2$$