CS 7180 / 7880 Lecture 3

Calibration and Decision Making

Decision Task

{rainy, not rainy}

{benign tumor, malign tumor}

Def. 1. (Decision Task). Fix an outcome set Y.

A decision task $T = (A, u)$.

A: action set {take umbrella, don't take}

{buy x shares of stock for $x \in \mathbb{Z}$}

{treatments for tumor}

$u: A \times Y \to \mathbb{R}$.

$u(a, y)$ utility of taking action a on outcome y.

Different people have different action set A and utility u.

E.g. Some people don't mind getting wet in rainy day.

When different people see the same prediction, they may take different actions.

How we translate predictions to actions.

If we know true outcome $y \in Y$, then we should take action a to maximize $u(a, y)$.

If we don't know y, but have a prediction p for y, then we take the "best response".

Def 2. (Best-Response Function). Suppose $Y = \{0, 1\}$.

Every decision task $T = (A, u)$ induces a best-response func

$BR_T : \underset{\text{predictions}}{[0, 1]} \longrightarrow A$ defined by

$$BR_T(p) = \arg\max_{a \in A} \mathbb{E}_{y \sim Ber(p)} [U(a, y)]$$

"trust" p as correct prediction

If p is correct, $BR_T(p)$ is a good action.

If p is wrong, $BR_T(p)$ may not be good.

Goal: evaluate predictions. Tool: scoring rules

Def. 3. (Scoring rules). Given pred. $p \in [0,1]$ outcome $y \in \{0,1\}$.

a scoring rule $S: [0,1] \times \{0,1\} \to \mathbb{R}$ gives

evaluation $S(p,y)$ for predicting p on outcome y.

Scoring rules induced by a decision task

Def. 4. Given decision task $T = (A, u)$.

We define an induced scoring rule $S_T : [0,1] \times \{0,1\} \to \mathbb{R}$ by

$$S_T(p, y) = u(BR_T(p), y) \quad \text{for every } p \in [0,1], y \in \{0,1\}$$

Example 1. $Y = \{0,1\}$, $A = \{0,1\}$. $T = (u, A)$

$$u(a, y) = \mathbb{I}(a = y).$$

$$BR_T(p) = \begin{cases} 0, & p < 0.5 \\ 1, & p \geq 0.5 \end{cases}$$

$$S_T(p, y) = u(BR_T(p), y) = \begin{cases} 1 & \text{if } p < 0.5, y = 0 \\ 0 & \text{if } p < 0.5, y = 1 \\ 0 & \text{if } p \geq 0.5, y = 0 \\ 1 & \text{if } p \geq 0.5, y = 1. \end{cases}$$

Example 2. $Y = \{0,1\}$, $A = \{0,1\}$.

$$U(a,y) = \begin{cases} 1, & \text{if } a = y \\ r, & \text{if } a=0, \ y=1 \\ 1-r, & \text{if } a=1, \ y=0. \end{cases}$$

$$BR_T(p) = \underset{a \in A = \{0,1\}}{\operatorname{argmax}} \ \underset{y \sim Ber(p)}{\mathbb{E}} [U(a,y)]$$

When $a=0$, $\mathbb{E}[U(a,y)] = p \cdot U(a,1) + (1-p) U(a,0)$

$$= p \cdot r + (1-p) \cdot 1$$

$$= 1 - p + pr$$

When $a=1$, $\mathbb{E}[U(a,y)] = p \cdot 1 + (1-p) \cdot (1-r)$

$$= p + 1 - p - r + pr$$

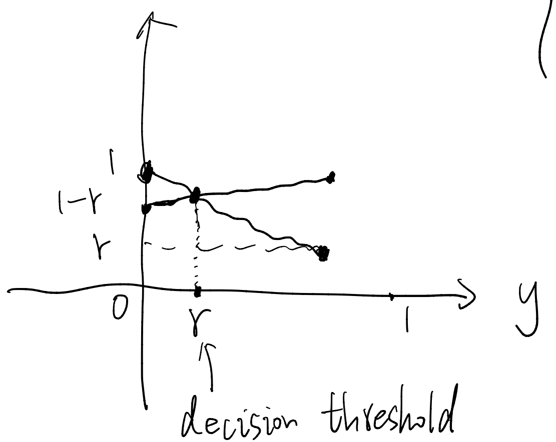
$$= 1 - r + pr$$

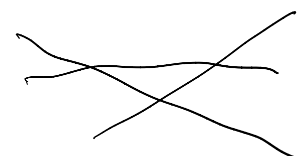
When $p < r$, $a = 0$ has higher expected utility.

$$BR_T(p) = 0$$

When $p \geq r$, $BR_T(p) = 1$

$$S_T(p, y) = u(BR_T(p), y) = \begin{cases} 1, & \text{if } p < r,\ y = 0 \\ r, & \text{if } p < r,\ y = 1 \\ 1-r, & \text{if } p \geq r,\ y = 0 \\ 1, & \text{if } p \geq r,\ y = 1 \end{cases}$$





Proper Scoring Rules.

Def. 5 (Proper Scoring Rules). We say S is proper if it "incentivizes" true prediction of $\mathbb{E}[y]$.

That is, For every $p, p' \in [0,1]$,

truth arbitrary

$$\mathop{\mathbb{E}}_{y \sim Ber(p)} [S(p, y)] \geq \mathop{\mathbb{E}}_{y \sim Ber(p)} [S(p', y)]$$

(Revelation Principle)

Thm. S is proper if and only if $S = S_T$ for some decision task T.

Proof. "if": $\mathop{\mathbb{E}}_{y \sim Ber(p)} [S_T(p, y)]$

$$= \mathop{\mathbb{E}}_{y\sim Ber(p)} [u(BR_T(p), y)]$$

$$\geq \mathop{\mathbb{E}}_{y\sim Ber(p)} [u(a, y)] \qquad \forall a \in A.$$

Take $a = BR_T(p')$.

$$\mathop{\mathbb{E}}_{y\sim Ber(p)} [S_T(p, y)] \geq \mathop{\mathbb{E}}_{y\sim Ber(p)} [u(BR_T(p'), y)] = \mathbb{E}[S_T(p', y)].$$

"only if": Take $A = [0, 1]$, $u(a, y) = s(a, y)$

$u(p, y) = s(p, y)$

By properness,

$$\mathop{\mathbb{E}}_{y\sim Ber(p)} [u(p, y)]$$

$$= \mathop{\mathbb{E}}_{y\sim Ber(p)} [s(p, y)]$$

$$\geq \mathbb{E} [s(p', y)]$$

$= \mathbb{E}[\, u(p', y) \,].$

$BR_T(p) = p.$

$$s(p, y) = u(BR_T(p), y)$$

In particular, for every $r \in [0,1]$ S_T is proper, where

$$S_T(p, y) = \begin{cases} 1, & \text{if } p < r,\ y = 0 \\ r, & \text{if } p < r,\ y = 1 \\ 1 - r, & \text{if } p \geq r,\ y = 0 \\ 1, & \text{if } p \geq r,\ y = 1 \end{cases}$$

Other proper scoring rules:

Brier score: $s(p, y) = 1 - (p - y)^2$

Cross entropy score: $s(p, y) = y \log p + (1-y) \cdot \log(1-p).$

Thm. Linear combination of proper scoring rules is proper.

Remark. There is a one-to-one correspondence between conv. functions $g:[0,1]\to\mathbb{R}$

and proper scoring rules.

Sequential Decision Making. Fix $T=(A,u)$.

Actions $a_1, \ldots, a_T \in A$

Outcomes $y_1, \ldots, y_T \in \{0,1\}$.

"Accuracy"

Average utility $\frac{1}{T}\sum_{t=1}^{T} u(a_t, y_t)$.

"normalization"

"indicator of intrinsic difficulty"

① (External) Regret: $-\left(\frac{1}{T}\sum_{t=1}^{T} u(a_t, y_t) - \max_{a\in A} \frac{1}{T}\sum_{t=1}^{T} u(a, y_t)\right)$

take same action a for all days t.

Regret can be positive or negative

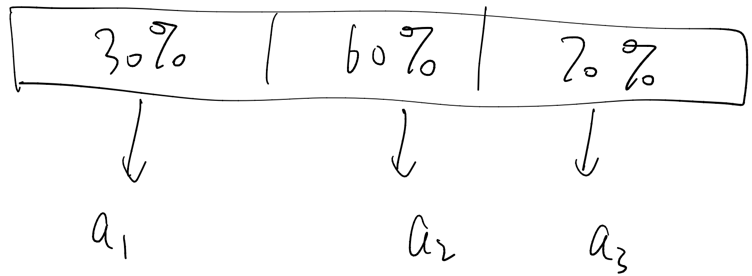
2 Swap regret. Always non-negative.

$$SR: -\left(\frac{1}{T} \sum_{t=1}^{T} u(a_t, y_t) - \max_{\sigma: A \to A} \frac{1}{T} \sum_{t=1}^{T} u(\sigma(a_t), y_t) \right)$$

"calibration"

$$E[y|p] = p$$

Perfect calibration $\Longleftrightarrow$ No swap regret for all decision tasks



Thus we can measure calibration error by swap regret of decision task.

Calibration Decision Loss [HW, FOCS 2024].