Comparación de Modelos: Costo-Beneficio

Matriz de Confusión

Datos reales

Predicciones del Modelo

		True default status		
		No	Yes	Total
	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Valor Esperado

$$E(f) = \sum_i p(x_i) \cdot f(x_i)$$

Si una decisión pudiera presentar *t* situaciones o salidas diferentes, con un resultado (positivo o negativo) asociado y una probabilidad asociado a cada uno de ellas, el valor esperado de esa decisión sería :

$$\boxed{\textit{Valor esperado} = p(s_1)\, v(s_1) + p(s_2) v(s_2) + \ldots + p(s_t) v(s_t)}$$

Veamos un ejemplo: Realizaremos una campaña de marketing en base a un dataset con información de nuestros clientes y campañas previas. El beneficio que obtendremos si un cliente responde es de $\mathbf{v}_R$ = \$99 pues compra el producto que le enviamos en la promoción. El costo si no responde a la campaña es 1, o sea $\mathbf{v}_{NR}$ = -\$1

Predicted \ Actual	p	n
Y	b(Y,p)	c(Y,n)
N	c(N,p)	b(N,n)

Matriz de confusión

	p	n
Y	56	7
N	5	42

Matriz de Beneficio

$$\begin{array}{cc} & \text{Actual} \\ \text{Predicted} & \begin{array}{c} \\ \mathbf{Y} \\ \mathbf{N} \end{array} \begin{pmatrix} \mathbf{p} & \mathbf{n} \\ 99 & -1 \\ 0 & 0 \end{pmatrix} \end{array}$$

$$\textit{Expected profit} = \overbrace{p(\mathbf{Y},\mathbf{p})\cdot b(\mathbf{Y},\mathbf{p})}^{\text{VP}} + \overbrace{p(\mathbf{N},\mathbf{p})\cdot b(\mathbf{N},\mathbf{p})}^{\text{FN}} + \overbrace{p(\mathbf{N},\mathbf{n})\cdot b(\mathbf{N},\mathbf{n})}^{\text{VN}} + \overbrace{p(\mathbf{Y},\mathbf{n})\cdot b(\mathbf{Y},\mathbf{n})}^{\text{FP}} \quad \textbf{(1)}$$

Por teoría de la probabilidad sabemos que: $p(x, y) = p(y)\cdot p(x \mid y)$

Por lo que podemos expresar a $p(\mathbf{Y},\mathbf{p})$ como $p(\mathbf{Y},\mathbf{p}) = p(\mathbf{Y} \mid \mathbf{p})\cdot p(\mathbf{p})$
Quedándonos el beneficio esperado:

$$\begin{aligned}\text{Expected profit} = \; & p(\mathbf{Y} \mid \mathbf{p})\cdot p(\mathbf{p})\cdot b(\mathbf{Y},\mathbf{p}) + p(\mathbf{N} \mid \mathbf{p})\cdot p(\mathbf{p})\cdot b(\mathbf{N},\mathbf{p}) + \\ & p(\mathbf{N} \mid \mathbf{n})\cdot p(\mathbf{n})\cdot b(\mathbf{N},\mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n})\cdot p(\mathbf{n})\cdot b(\mathbf{Y},\mathbf{n})\end{aligned}$$

$$\boxed{\begin{aligned}\text{Expected profit} = \; & p(\mathbf{p}) \cdot [p(\mathbf{Y} \mid \mathbf{p})\cdot b(\mathbf{Y},\mathbf{p}) + p(\mathbf{N} \mid \mathbf{p})\cdot c(\mathbf{N},\mathbf{p})] + \\ & p(\mathbf{n}) \cdot [p(\mathbf{N} \mid \mathbf{n})\cdot b(\mathbf{N},\mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n})\cdot c(\mathbf{Y},\mathbf{n})]\end{aligned}} \quad \textbf{(2)}$$

Matriz de confusión

	p	n
Y	56	7
N	5	42

$T = 110$

$P = 61$ $\quad N = 49$

$p(\mathbf{p}) = 0.55$ $\quad p(\mathbf{n}) = 0.45$

$tp\ rate = 56/61 = 0.92$ $\quad fp\ rate = 7/49 = 0.14$

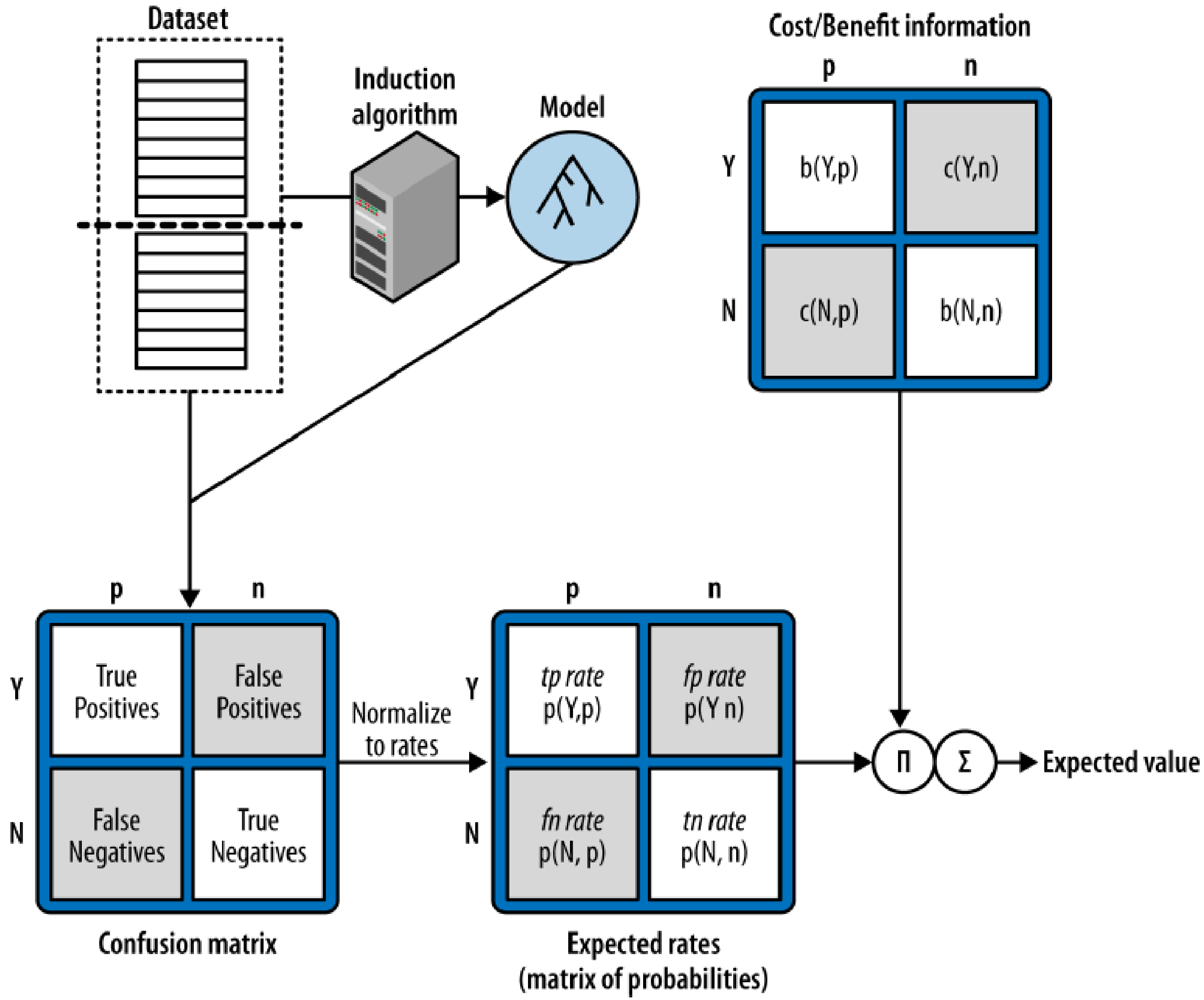
$fn\ rate = 5/61 = 0.08$ $\quad tn\ rate = 42/49 = 0.86$

Matriz de Costo/Beneficio

$$\text{Predicted} \begin{array}{c} \\ \mathbf{Y} \\ \mathbf{N} \end{array} \overset{\text{Actual}}{\begin{pmatrix} \mathbf{p} & \mathbf{n} \\ 99 & -1 \\ 0 & 0 \end{pmatrix}}$$

Reemplazando estos valores en (2), obtenemos:

$$\begin{aligned}
\text{expected profit} &= p(\mathbf{p}) \cdot [p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + \\
&\quad\ p(\mathbf{n}) \cdot [p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{p}) \cdot c(\mathbf{Y}, \mathbf{n})] \\
&= 0.55 \cdot [0.92 \cdot b(\mathbf{Y}, \mathbf{p}) + 0.08 \cdot b(\mathbf{N}, \mathbf{p})] + \\
&\quad\ 0.45 \cdot [0.86 \cdot b(\mathbf{N}, \mathbf{n}) + 0.14 \cdot p(\mathbf{Y}, \mathbf{n})] \\
&= 0.55 \cdot [0.92 \cdot 99 + 0.08 \cdot 0] + \\
&\quad\ 0.45 \cdot [0.86 \cdot 0 + 0.14 \cdot\ -1] \\
&= 50.1 - 0.063 \\
&\approx \mathbf{\$50.04}
\end{aligned}$$


Dataset
Induction algorithm
Model
Cost/Benefit information
p
n
Y
b(Y,p)
c(Y,n)
N
c(N,p)
b(N,n)
p
n
Y
True Positives
False Positives
N
False Negatives
True Negatives
Normalize to rates
p
n
Y
tp rate p(Y,p)
fp rate p(Y n)
N
fn rate p(N, p)
tn rate p(N, n)
Π
Σ
Expected value
Confusion matrix
Expected rates (matrix of probabilities)