ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Introducción a los modelos de clasificación y a su evaluación

Andrés G. Abad, Ph.D.

Agenda

Introducción al problema de clasificación

Evaluando la precisión de la predicción

Validación cruzada

Matriz de confusión

Curva ROC (receiver operating characteristic)

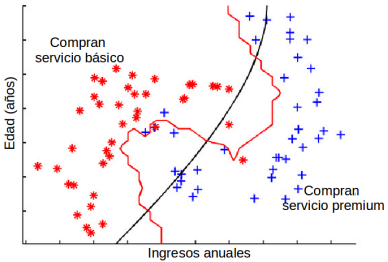
Definición del problema de clasificación I

- Un objeto $\mathbf{x} = [x_1, \ldots, x_p]$, con características x_i, pertenece exactamente a una clases $c \in \{1, 2, \ldots, C\}$.
- Asumimos que tenemos un conjunto de datos

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, c^{(1)}), \ldots, (\mathbf{x}^{(n)}, c^{(n)})\}$$

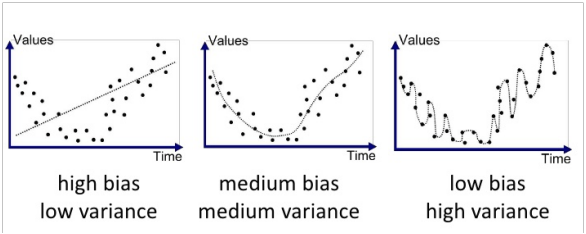
- Buscamos una función $\hat{f}$ que asigne $\mathbf{x}^{(i)}$ a $c^{(i)}$ lo mejor posible:

$$\hat{f} = \arg\min_{f} \mathbb{P}_{(\mathbf{x},c)}[\mathbb{1}(f(\mathbf{x}) \neq c)]$$



- Objeto $\mathbf{x}$ pertenece a una de dos clases: $\{Basico, Premium\}$
- Objeto $\mathbf{x}$ medidos en dos características: x_1 ingresos anuales, y x_2 edad en años
- Dos clasificadores $\hat{f}$'s: convexo-cuadrático (linea negra) y no-convexo (linea roja)

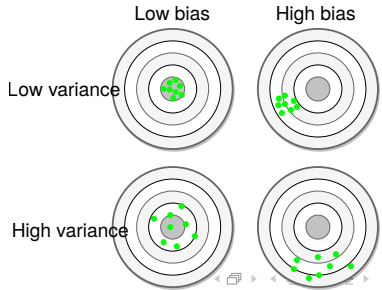
El tradeoff entre Sesgo y Variance I



$$\mathbb{E}\left[\left(y-\hat{f}(x)\right)^2\right] = \left(\text{Bias}\left[\hat{f}(x)\right]\right)^2 + \text{Var}\left[\hat{f}(x)\right] + \sigma^2$$

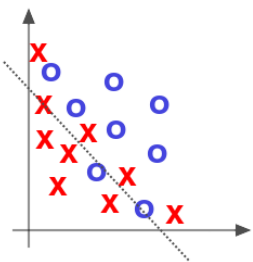
$$\text{Bias}\left[\hat{f}(x)\right] = \mathbb{E}\left[\hat{f}(x) - f(x)\right]$$

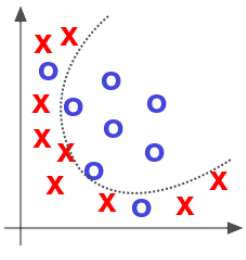
$$\text{Var}\left[\hat{f}(x)\right] = \mathbb{E}\left[\hat{f}(x)^2\right] - \left(\mathbb{E}\left[\hat{f}(x)\right]\right)^2$$

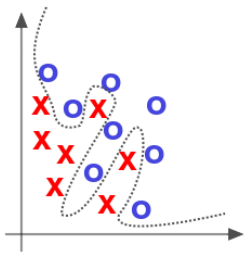


El tradeoff entre Sesgo y Variance II

Clasificación





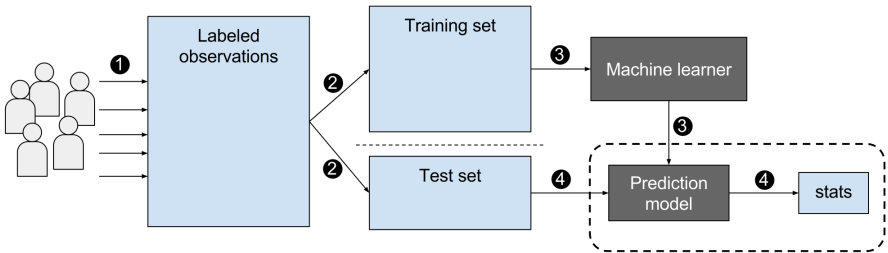


Evaluando desempeño del modelo I

Para evaluar el desempeño de un modelo generalmente se utilizan algunas de las dos siguientes medidas

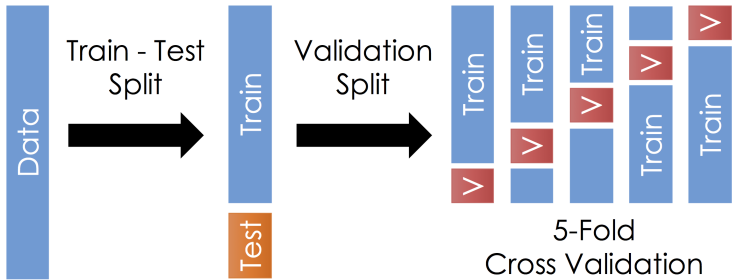
- **Modelo explicativo**: R^2 (o considerando complejidad del modelo R^2_{adj})
- **Modelo predictivo**: RSS

Para estimar el RSS necesitamos particionar los datos en: (1) datos de entrenamiento y (2) datos de prueba



Validación cruzada (cross validation) I

Para obtener estimaciones de la distribución de los estimadores (e.g. precisión de la predicción) podemos utilizar la validación cruzada (cross validation)



Esto nos permite tener estimaciones de intervalos (como intervalos de confianza)

Evaluando desempeño en clasificación binaria I

		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive / Σ Test outcome positive	False discovery rate (FDR) = Σ False positive / Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = Σ False negative / Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative / Σ Test outcome negative
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = Σ True positive / Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive / Σ Condition negative	Accuracy (ACC) = Σ True positive + Σ True negative / Σ Total population	
	Negative likelihood ratio (**LR−**) = FNR/TNR	False negative rate (FNR) = Σ False negative / Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative / Σ Condition negative		
	Diagnostic odds ratio (**DOR**) = LR+/LR−				

Matriz de confusión, accuracy y coeficiente (cohen's) kappa I

		Real		
		Positivo	Negativo	Total
Predicción	Positivo	a	b	$a+b$
	Negativo	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Tenemos que el Accuracy (ACC)

$$ACC = \frac{a+d}{N}$$

$$\text{Error} = 1 - ACC$$

Matriz de confusión, accuracy y coeficiente (cohen's) kappa II

El coeficiente (cohen's) kappa κ mide el acuerdo entre dos fuentes donde cada una clasifica N items en C clases mutuamente excluyentes.

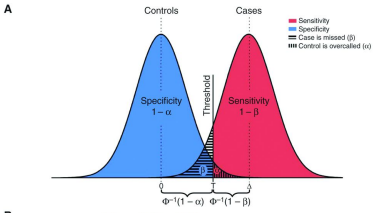
$$\kappa \equiv \frac{ACC - p_e}{1 - p_e},$$

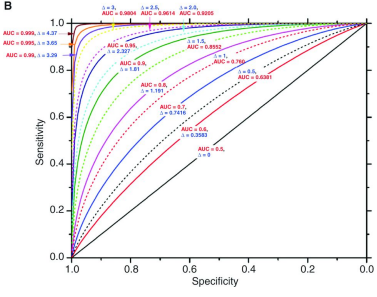
donde p_e es la probabilidad hipotética de acuerdos al azar.

		A: Si	A: No	Total
B	Si	a	b	$a+b$
B	No	c	d	$c+d$
	Total	$a+c$	$b+d$	N

- $ACC = \frac{a+d}{N}$
- $p_{Si} = \frac{a+b}{N} \cdot \frac{a+c}{N}$
- $p_{No} = \frac{c+d}{N} \cdot \frac{b+d}{N}$
- $p_e = p_{Si} + p_{No}$

Curva ROC (receiver operating characteristic) I





		Real		
		Positivo	Negativo	Total
Predicción	Positivo	a	b	$a+b$
	Negativo	c	d	$c+d$
	Total	$a+c$	$b+d$	N

- Specificity $= \frac{d}{b+d}$
- Sensitivity $= \frac{a}{a+c}$