Modelos Lineales con R: Unidad 4

Andrés García Medina
Facultad de Ciencias, Universidad Autónoma de Baja California
https://sites.google.com/view/andresgm/home

Enero-Junio, 2024

1. Regresión Logística en la práctica

Uso de los modelos

- Los modelos de regresión logística contribuyen a:
  - Identificar factores de riesgo.
  - Evaluar el riego (probabilidad de ocurrencia del evento respuesta) para individuos específicos.
  - Clasificar / discriminar a grupos de individuos como alto riesgo o bajo riesgo.
- La formulación y estimación del modelo

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-x^T \beta}}$$

nos da una fórmula que nos permite evaluar el riesgo de un individuo específico (i.e. para un vector, x, de atributos específicos).

Problemas de clasificación

- Clasificamos a un individuo como de alto riesgo si la probabilidad de ocurrencia es mayor o igual a cierto umbral; entonces, diremos que un individuo con atributos x es de alto riesgo si

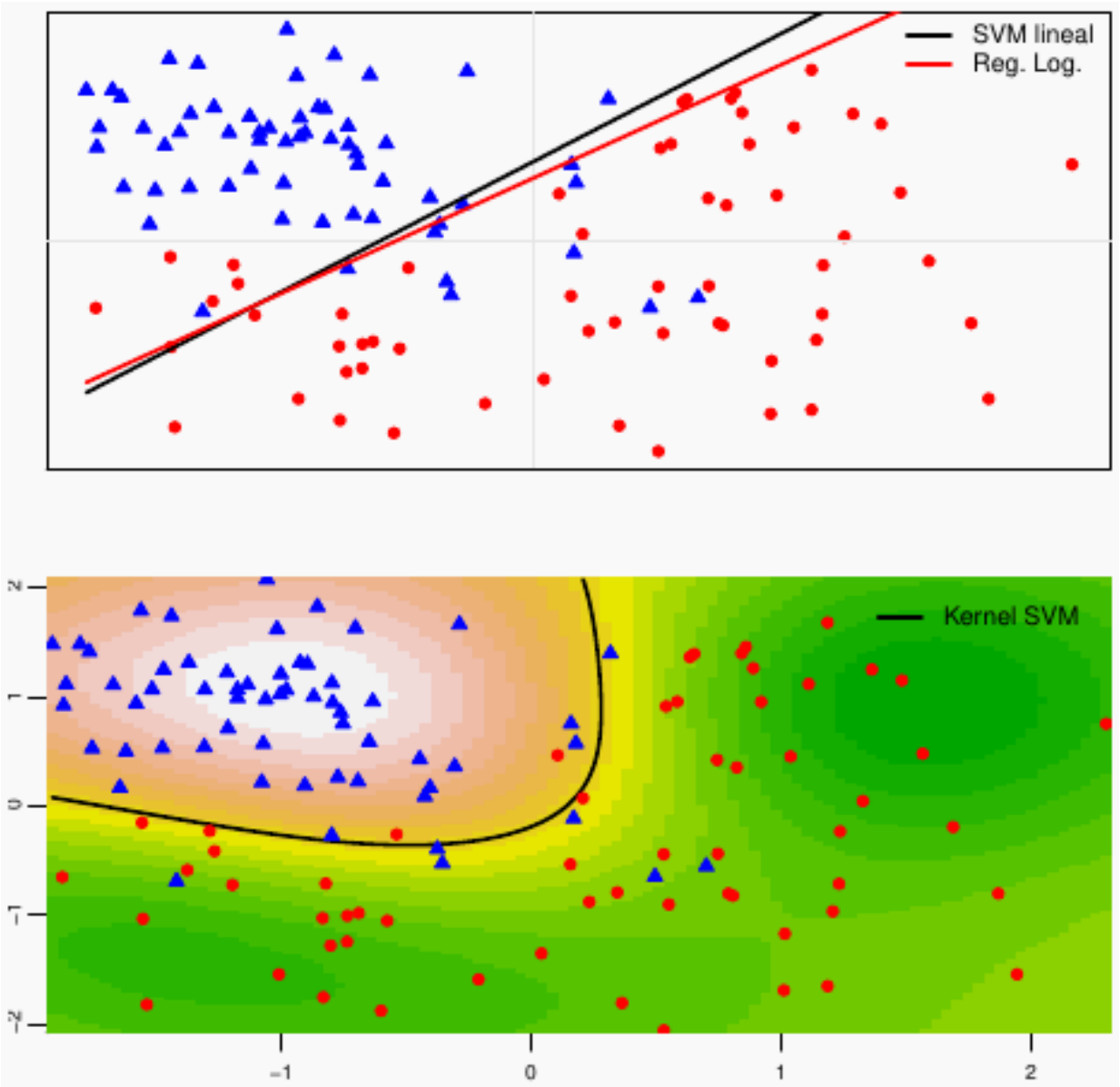
$$\frac{1}{1+e^{-x^T\beta}} > p$$

esto es, si

$$x^T\beta > \log\frac{p}{1-p}$$

- En el caso particular de $p = 0.5$ tenemos que los individuos de alto riesgo satisfacen $x^T\beta > 0$.
- La gráfica siguiente muestra el comportamiento de regresión logística, como función discriminante, comparada con el clasificador SVM.
- Regresión logística, en su forma estándar, no es tan poderosa como los SVM's, sin embargo, son muy útiles por sus usos en la identificación de factores de riesgo, lo cual, no es inmediato con los SVM's kernelizados.

Regresión logística vs SVM's



Curva ROC

- Cuando usamos un modelo de regresión logísica para clasificación, tenemos que definir el umbral, p, a partir del cual declaramos un "positivo".
- Las curvas ROC grafican las tasas TPR vs FPR para diferentes umbrales p.

$$TPR = \text{True Positive Rate} = \frac{TP}{P} = \text{"sensitividad"}$$

$$FPR = \text{False Positive Rate} = \frac{FP}{N} = 1 - \text{"especificidad"}$$

	Observados	
	1	0
Decisión = 1	TP	FP
Decisión = 0	FN	TN
	P	N

- **Error Tipo I**: Rechazo de la hipótesis nula cuando ésta es verdadera.
- **Error Tipo II**: No rechazo de la hipótesis nula cuando esta es falsa.

Curva ROC

- La *Sesitividad* y *especificidad* recaen en un solo punto de corte para clasificar un resultado como positivo.
- Una descripción más completa es a través del area bajo la curva ROC (Receiver Operating Characteristic).
- La idea original es detectar una señal en presencia de ruido.
- ROC muestra la probabilidad de detectar una señal verdadera (sensitividad) respecto a una señal falsa (1 - especificidad) para diferentes valores de cortes.
- El área bajo la curva ROC, denominada AUC nos da una medida de la habilida del modelo para discriminar entre los sujetos que presentan la caracteristica o no (enfermedad coronaria).

ROC con librería ROCR