La tarea de clasificación en detalle (I)

Alfredo Cuesta Infante

E. T. S. Ingeniería Informática
Universidad Rey Juan Carlos

Master Univ. en Visión Artificial
Reconocimiento de Patrones

Contenidos

Contenidos

Regresión lineal

¿Podemos clasificar con regresión?

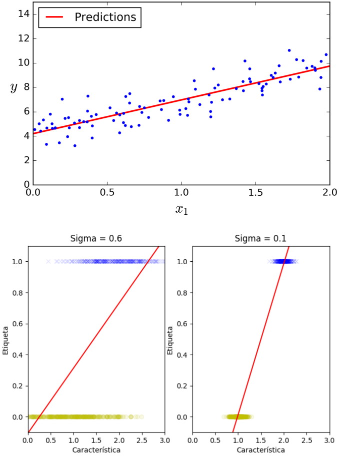

Figura: (Arriba.) Regresión lineal con 1 sola dimensión. (Abajo.) ¿Tiene sentido la regresión lineal para hacer clasificación binaria? [Fuente: Original de A. Cuesta]

Clasificación con Regresión Lineal

Función discriminante

- La regresión devuelve un valor continuo $\hat{y}$ para un ejemplo x dado.
- Pero en clasificación buscamos una etiqueta $\hat{t}$, que es discreta.
- Con la **función discriminante** obtenemos $\hat{t}$ a partir de $\hat{y}$.

Ej.:

$$\hat{t} = \begin{cases} \text{“etiqueta 0”} & \text{si} \quad \hat{y} < \theta \\ \text{“etiqueta 1”} & \text{si} \quad \hat{y} \geq \theta \end{cases}$$

- En el caso (muy concreto) de regresión lineal con ejemplos de una sóla dimensión, como en la figura anterior:

$$\hat{y} = w_0 + w_1 x_1 = [w_0, w_1] \begin{bmatrix} 1 \\ x_1 \end{bmatrix} = \mathbf{w}^T \mathbf{x}$$

- θ es el umbral (*threshold)* con el que discriminamos.

Clasificación con Regresión Lineal

Clasificar es variar el umbral

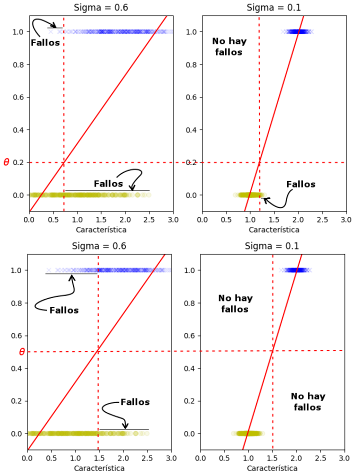

Figura: Variando el umbral logramos separar mejor las dos clases, aunque el conjunto con σ =0.6 no es linealmente separable porque las dos clases se solapan. [Fuente: Original de A. Cuesta]

Regresión logística

Función sigmoide

- Transforma $\hat{y}$ en una probabilidad $\hat{p}$

$$\hat{p} = S(\hat{y}) = S\left(\mathbf{w}^T\mathbf{x}\right) = \frac{1}{1 + \exp\left(-\mathbf{w}^T\mathbf{x}\right)}$$

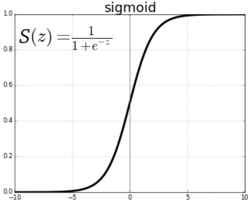

Figura: Función sigmoide. [Fuente: Anónimo @ Internet]

Función discriminante

$$\hat{t} = \begin{cases} \text{“etiqueta 0”} & \text{si} \quad \hat{p} < \theta \\ \text{“etiqueta 1”} & \text{si} \quad \hat{p} \geq \theta. \end{cases}$$

Notación n-dimensional

Cuadro: Tabla resumen de clasificación mediante regresión.

Reg. Lineal	Reg. Logística
Conjunto de m ejemplos n dimensionales etiquetados	
$\mathbf{w} = [w_0, w_1, w_2, \ldots, w_n]^T$	
$\mathbf{x}^{(i)} = [1, x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]^T$	
$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)T} \\ \vdots \\ \mathbf{x}^{(m)T} \end{bmatrix} = \begin{bmatrix} 1, & x_1^{(1)}, & x_2^{(1)}, & \ldots & x_n^{(1)} \\ 1, & x_1^{(2)}, & x_2^{(2)}, & \ldots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1, & x_1^{(m)}, & x_2^{(m)}, & \ldots & x_n^{(m)} \end{bmatrix}$	
$\hat{y}^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)}$	$\hat{p}^{(i)} = 1/(1 + \exp(-\mathbf{w}^T \mathbf{x}^{(i)}))$
$\hat{t}^{(i)} = \begin{cases} \text{“etiqueta 0”} & \text{si } \hat{y} < \theta \\ \text{“etiqueta 1”} & \text{si } \hat{y} \geq \theta \end{cases}$	$\hat{t}^{(i)} = \begin{cases} \text{“etiqueta 0”} & \text{si } \hat{p}^{(i)} < \theta \\ \text{“etiqueta 1”} & \text{si } \hat{p}^{(i)} \geq \theta \end{cases}$

Función de coste

Búsqueda del mejor

- Datos : $\mathbf{X}$ y $\mathbf{t}$
- Modelo : Lineal
- Parámetros : Aquella combinación **mejor** para los datos dados
- Hace falta **medir** = una **función de coste**

$$J(\mathbf{w}; \mathbf{X}, \mathbf{t})$$

$\Rightarrow$ El **mejor** = el argumento que la **minimiza**

$$\mathbf{w}^* = \arg\min_{w} \left(J(\mathbf{w}; \mathbf{X}, \mathbf{t}) \right).$$

- Funciones de coste típicas de regresión lineal:

  MSE, MAE, RMSE

- Funciones de coste típicas de regresión logística:

  La función *log-loss*.

Método de optimización

Fórmula cerrada

- En el caso de la regresión lineal (y alguno más) se puede obtener una fórmula cerrada para calcular $\mathbf{w}^*$
- Es tan sencillo como resolver el siguiente sistema lineal:

$$\begin{bmatrix} 1, & x_1^{(1)}, & x_2^{(1)}, & \dots & x_n^{(1)} \\ 1, & x_1^{(2)}, & x_2^{(2)}, & \dots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1, & x_1^{(m)}, & x_2^{(m)}, & \dots & x_n^{(m)} \end{bmatrix} \begin{bmatrix} w_0^* \\ w_1^* \\ w_2^* \\ \vdots \\ w_n^* \end{bmatrix} = \begin{bmatrix} t^{(1)} \\ t^{(2)} \\ \vdots \\ t^{(m)} \end{bmatrix};$$

Es decir $\mathbf{X}\mathbf{w}^* = \mathbf{t}$
Despejando $\mathbf{w}^*$ tenemos que:

$$\mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t}$$

- Otro modo es resolver:

$$\frac{\partial J(\mathbf{w}, \mathbf{X}, \mathbf{t})}{\partial w_j} = 0, \quad \text{con} \quad j = 1, 2, \dots, n.$$

Pero es preferible acudir a una solución numérica

Método de optimización numérico

Descenso del gradiente

- Se alcanza $\mathbf{w}^*$ mediante aproximaciones sucesivas.
- Parece lógico intentar llegar por el *camino más corto*.
- Es decir, el que baja más rápido.
- La dirección de mayor descenso es $-\nabla J$.
- **Regla de actualización de pesos**

$$\mathbf{w}' = \mathbf{w} - \eta \nabla J(\mathbf{w})$$

donde

- $\mathbf{w}'$ es el nuevo vector de pesos
- $\mathbf{w}$ es el vector de pesos del que partimos
- η es el paso (ratio) de aprendizaje

Método de optimización numérico

Descenso del gradiente

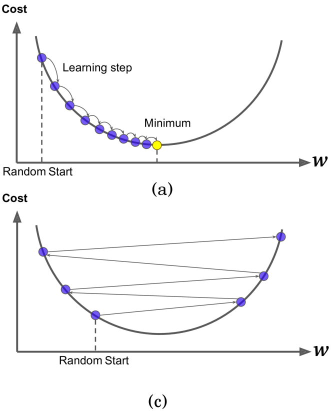

Figura: Efecto del ratio de aprendizaje η en el descenso del gradiente para $f(x)$. (a) Con un buen ratio, (b) con un ratio demasiado pequeño, (c) con uno demasiado grande, y (d) cuando J tiene varios mínimos locales y mesetas (*plateau*).

Criterio de parada

- Límite en el número de iteraciones
- Convergencia a un vector de pesos
- *Early Stopping*

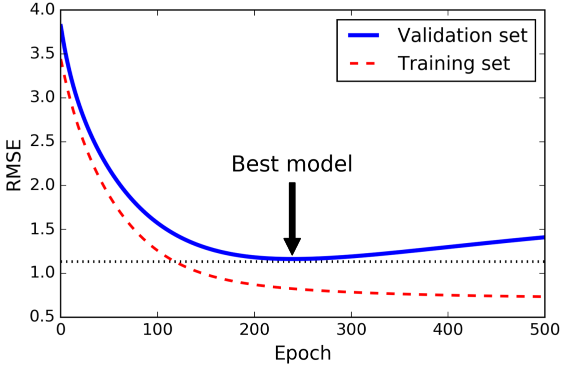

Figura: Utilización del conjunto de validación para decidir cuando debemos parar de aprender.

Regularización

Añadir un término en la función de coste para penalizar los pesos

- Ridge

$$J_{\mathrm{R}}(\mathbf{w};\mathbf{X},\mathbf{t}) = J(\mathbf{w};\mathbf{X},\mathbf{t}) + \frac{\alpha}{2}\sum_{j=1}^{n} w_j^2, \quad \text{con} \quad \alpha \geq 0.$$

- Lasso

$$J_{\mathrm{L}}(\mathbf{w};\mathbf{X},\mathbf{t}) = J(\mathbf{w};\mathbf{X},\mathbf{t}) + \alpha\sum_{j=1}^{n} |w_j|, \quad \text{con} \quad \alpha \geq 0.$$

- Elastic Net

$$J_{\mathrm{EN}}(\mathbf{w};\mathbf{X},\mathbf{t}) = J(\mathbf{w};\mathbf{X},\mathbf{t}) + (r)J_{\mathrm{L}}(\mathbf{w};\mathbf{X},\mathbf{t}) + (1-r)J_{\mathrm{R}}(\mathbf{w};\mathbf{X},\mathbf{t}),$$

$$\text{con} \quad 0 \leq r \leq 1.$$

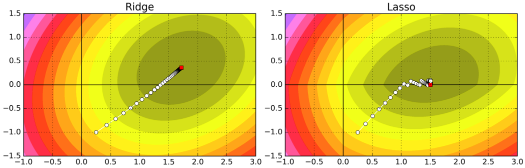

Figura: Descenso del gradiente en 2 dims. con regularización *Ridge* y *Lasso*

Contenidos

Formulación del problema de optimización
Regresión lineal
Regresión logística
Notación n-dimensional
Función de coste
Método de optimización
Criterio de parada
Regularización
Ajuste fino de modelos
Medidas de bondad
Tabla de Confusión
Precision, Recall, F-score
Curva ROC

Ajuste fino de modelos

Parámetros vs. Hiperparámetros

- Los hiperparámetros configuran el proceso de aprendizaje
- Los parámetros pertenecen al modelo

Métodos de búsqueda

- Ordenada (*grid*)
- Aleatoria
- Heurística
- Métodos de ensamblado (combinación de clasificadores)

Contenidos

Medidas de bondad

Tabla de Confusión

		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–				

Figura: Tabla de confusión en clasificación binaria y medidas derivadas. [Fuente: Wikipedia]

Medidas de bondad

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

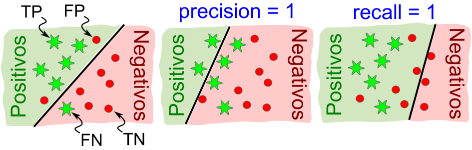

Figura: Un clasificador puede lograr un *precision* o un *recall* del 100 %. Basta con que no haya ningún FP o ningún FN respectivamente. [Fuente: Original de A. Cuesta]

Medidas de bondad

Compromiso Precision–Recall

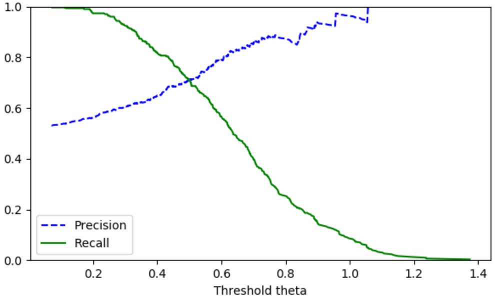

Figura: Curvas de *precision* y *recall* para el conjunto de datos de la Figura 1 con $\sigma = 0{,}6$. [Fuente: Original de A. Cuesta]

F-score

$$\mathrm{F-score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Medidas de bondad

Curva ROC

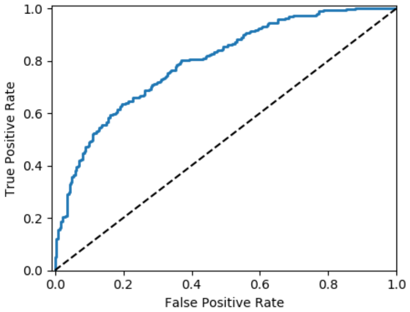

Figura: Curvas ROC para el conjunto de datos de la Figura 1 con $\sigma = 0{,}6$. [Fuente: Original de A. Cuesta]