

Notas sobre Métricas, Escalado de Datos y Redes Neuronales

Autor: Daniel Garbanzo c.2022117129
Escuela de Ingeniería en Computación
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica

Abstract—Este documento resume conceptos clave acerca de la evaluación de modelos predictivos, técnicas de escalado de características (normalización y estandarización) y fundamentos en redes neuronales. Se discuten métricas de clasificación, la importancia de un adecuado escalado de *features* y se presentan casos de estudio que ilustran aplicaciones prácticas en problemas de clasificación.

I. INTRODUCCIÓN

La correcta evaluación de modelos predictivos es esencial para determinar el rendimiento y la eficacia de algoritmos en problemas de clasificación. Este documento recopila conceptos vistos en clase, abarcando desde métricas de evaluación hasta técnicas de normalización y estandarización, finalizando con una introducción a redes neuronales y la clasificación de datos, tal como se abordó en la sesión de la Semana 7.

II. MÉTRICAS DE EVALUACIÓN DE MODELOS

A. Matriz de Confusión

Actual value	Positive	TP	FN
	Negative	FP	TN
		Positive	Negative
		Predicted value	

Figura 1. Ejemplo de matriz de confusión.

La matriz de confusión se utiliza en problemas de clasificación binaria y permite distinguir entre:

- **Positivo Verdadero:** El modelo predice la clase positiva correctamente (por ejemplo, un doctor informando a una mujer que está embarazada cuando efectivamente lo está).
- **Falso Positivo (Error Tipo I):** El modelo predice incorrectamente la clase positiva (por ejemplo, un doctor informando a un adulto mayor que está embarazado).
- **Negativo Verdadero:** El modelo predice la clase negativa de forma correcta.

- **Falso Negativo (Error Tipo II):** El modelo falla al no detectar una instancia positiva (por ejemplo, un doctor informando a una mujer embarazada que no lo está).

B. Otras Métricas

A continuación se describen algunas de las métricas utilizadas para evaluar modelos de clasificación:

- a. **Accuracy:** Se define como la proporción de clasificaciones correctas respecto al total de intentos.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Figura 2. Fórmula para el cálculo de Accuracy.

- b. **Precision:** Mide la tasa de verdaderos positivos entre todas las predicciones positivas, evaluando los errores de tipo I.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Figura 3. Fórmula para el cálculo de Precision.

- c. **Recall:** Evalúa la proporción de verdaderos positivos detectados, considerando los errores de tipo II.

- d. **F1-Score:** Es la media armónica entre Precision y Recall, resultando particularmente útil en escenarios con clases desbalanceadas.

C. ROC y AUC

El análisis mediante la curva ROC y su área bajo la curva (AUC) permite evaluar la capacidad del clasificador para distinguir entre clases. Un AUC menor o igual a

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Figura 4. Fórmula para el cálculo de Recall.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figura 5. Fórmula para el cálculo del F1-Score.

0.5 indica un rendimiento similar al de un clasificador aleatorio.

III. NORMALIZACIÓN Y ESTANDARIZACIÓN

El escalado de características es fundamental para preparar los datos antes del entrenamiento de los modelos. A continuación, se describen dos técnicas comunes:

A. Normalización

La normalización ajusta los datos para que se encuentren dentro de un rango específico, normalmente entre 0 y 1 (o [-1, 1]). La fórmula utilizada es:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Esto asegura que las diferentes características tengan la misma escala, lo que es particularmente importante cuando las unidades o rangos de los datos varían considerablemente.

B. Estandarización

La estandarización transforma los datos para que presenten una media de 0 y una desviación estándar de 1, de manera que se siga una distribución normal (Z-score):

$$x_{\text{std}} = \frac{x - \mu}{\sigma} \quad (2)$$

Esta técnica es preferible cuando los datos tienen distribución gaussiana o se desea mitigar el efecto de los outliers.

IV. REDES NEURONALES Y CLASIFICACIÓN

A. Caso de Estudio: Clasificación de MNIST

El conjunto de datos MNIST se utiliza para la clasificación de dígitos escritos a mano. En este caso:

- Las imágenes originales, por ejemplo de 128x128 píxeles, se transforman a un tamaño menor (28x28 píxeles), generando 784 *features* por imagen.
- El objetivo es clasificar correctamente dígitos del 0 al 9, utilizando técnicas como la regresión logística y las redes neuronales.

B. Procesamiento y Arquitectura

Para preparar la entrada de las redes neuronales y optimizar los cálculos se aplican las siguientes técnicas:

a. **Transformación (Flatten)**: Las imágenes se convierten en vectores unidimensionales, concatenando todas las filas para formar un único vector de 784 componentes.

b. **One-Hot Encoding**: Cada dígito se representa mediante un vector *one-hot* para indicar de forma exclusiva la clase detectada.

	0	1	2	3	4	5	6	7	8	9
0	1	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	1

Digit
Class

One hot encoded vector

Figura 6. Representación de un vector one-hot.

c. **Optimización y Cálculo de Pesos**: En lugar de calcular múltiples regresiones lineales por separado, se aprovechan las operaciones de álgebra lineal utilizando:

- Una matriz de pesos W donde cada columna representa una regresión logística.
- Un vector de sesgos b que se ajusta durante el entrenamiento.

El proceso de optimización se realiza mediante *forward propagation* y *back propagation*.

V. CONCLUSIONES

Se han presentado diversas métricas para evaluar modelos predictivos y se ha discutido la importancia del escalado de características mediante normalización y estandarización. Además, se han explicado conceptos fundamentales en redes neuronales aplicados a la clasificación, poniendo énfasis en la transformación de imágenes y el uso de múltiples regresiones logísticas. La correcta aplicación de estas técnicas es esencial para mejorar el rendimiento de los algoritmos en problemas reales.