

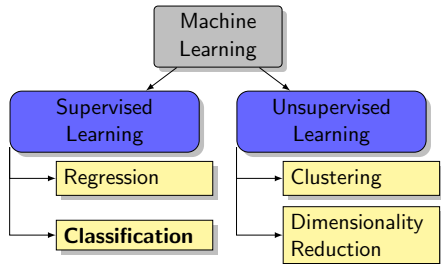
Clasificación

Dr. Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Diplomado Ciencia de Datos con Python

Table of Contents

- 1 Introducción: La tarea de clasificación
- 2 Métricas de desempeño
- 3 Comparación de algoritmos



La tarea de clasificación

¿Qué tienen en común las siguientes tareas?



La tarea de clasificación

¿Qué tienen en común las siguientes tareas?

Hello Friends! We hope you had a pleasant week. Last weeks trivia questions was:

What do these 3 films have in common: One Crazy Summer, Whispers in the Dark, Moby Dick?

Answer: Nantucket Island

No Spam

IMPORTANT INFORMATION:

The new domain names are finally available to the general public at discount prices. Now you can register one of the exciting new .BIZ or .INFO domain names, as well as the original .COM and .NET names for just \$14.95. These brand new domain extensions were recently approved by ICANN and have the same rights as the original .COM and .NET domain names. The biggest benefit is of-course that the .BIZ and .INFO domain names are currently more available. i.e. it will be much easier to register an attractive and easy-to-remember domain name for the same price. Visit: <http://www.affordable-domains.com> today for more info.

Spam

If you have an internal zip drive (not sure about external) and you bios supports using a zip as floppy drive, you could use a bootable zip disk with all the relevant dos utils.

No Spam

La tarea de clasificación

Clasificación

Problema supervisado en el cual el objetivo es asignar una etiqueta o categoría a cada ejemplo de un conjunto de datos.

La tarea de clasificación consiste en entrenar un modelo para predecir a qué clase pertenece una nueva observación. Esto lo hacemos basándonos en un conjunto de datos etiquetados donde las categorías son conocidas.

- **Clasificación Binaria:** Dos etiquetas, mutuamente exclusivas.

La tarea de clasificación

Clasificación

Problema supervisado en el cual el objetivo es asignar una etiqueta o categoría a cada ejemplo de un conjunto de datos.

La tarea de clasificación consiste en entrenar un modelo para predecir a qué clase pertenece una nueva observación. Esto lo hacemos basándonos en un conjunto de datos etiquetados donde las categorías son conocidas.

- **Clasificación Binaria:** Dos etiquetas, mutuamente exclusivas.
- **Clasificación Multi-clase:** Varias etiquetas mutuamente excluyentes.

La tarea de clasificación

Clasificación

Problema supervisado en el cual el objetivo es asignar una etiqueta o categoría a cada ejemplo de un conjunto de datos.

La tarea de clasificación consiste en entrenar un modelo para predecir a qué clase pertenece una nueva observación. Esto lo hacemos basándonos en un conjunto de datos etiquetados donde las categorías son conocidas.

- **Clasificación Binaria:** Dos etiquetas, mutuamente exclusivas.
- **Clasificación Multi-clase:** Varias etiquetas mutuamente excluyentes.
- **Clasificación Multi-etiqueta:** Cada instancia tiene varias etiquetas.

Clasificación Binaria

<p>Hello Friends! We hope you had a pleasant week. Last weeks trivia questions was:</p> <p>What do these 3 films have in common: One Crazy Summer, Whispers in the Dark, Moby Dick?</p> <p>Answer: Nantucket Island</p>	No Spam
<p>IMPORTANT INFORMATION:</p> <p>The new domain names are finally available to the general public at discount prices. Now you can register one of the exciting new .BIZ or .INFO domain names, as well as the original .COM and .NET names for just \$14.95. These brand new domain extensions were recently approved by ICANN and have the same rights as the original .COM and .NET domain names. The biggest benefit is of-course that the .BIZ and .INFO domain names are currently more available. i.e. it will be much easier to register an attractive and easy-to-remember domain name for the same price. Visit: http://www.affordable-domains.com today for more info.</p>	Spam
<p>If you have an internal zip drive (not sure about external) and you bios supports using a zip as floppy drive, you could use a bootable zip disk with all the relevant dos utils.</p>	No Spam

Clasificación Binaria

Texto	ℓ
Hello Friends! We hope you had a pleasant week. Last weeks trivia questions was ...	0
IMPORTANT INFORMATION: The new domain names are finally available to the general...	1
...	...

Clasificación Binaria

w_1	...	w_M	ℓ
2	...	0	1
...

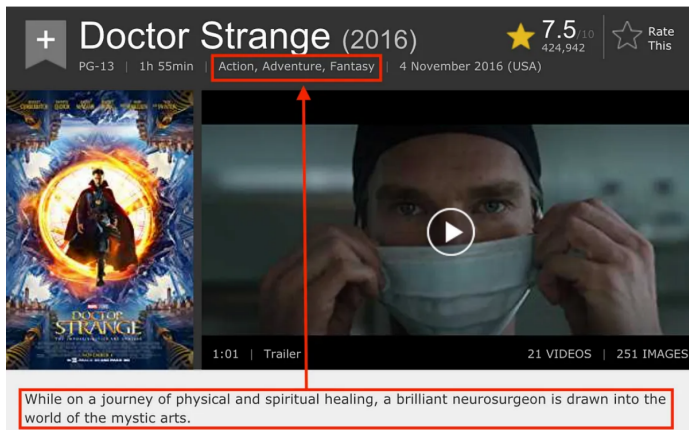
Clasificación Multi-clase



Clasificación Multi-clase

p_1	...	p_{784}	ℓ
0	...	57	0
...

Clasificación Multi-etiqueta



The image shows the movie page for "Doctor Strange (2016)" on a platform. The title "Doctor Strange (2016)" is at the top left, followed by the rating "7.5/10" and "424,942" reviews. The genre "Action, Adventure, Fantasy" is highlighted with a red box. Below the title, the movie poster and a video player are visible. The video player shows a scene of Doctor Strange putting on his mask. The description at the bottom is also highlighted with a red box: "While on a journey of physical and spiritual healing, a brilliant neurosurgeon is drawn into the world of the mystic arts." A red arrow points from the genre box to the description box.

Doctor Strange (2016) 7.5/10 424,942 Rate This

PG-13 | 1h 55min | **Action, Adventure, Fantasy** | 4 November 2016 (USA)

1:01 | Trailer 21 VIDEOS | 251 IMAGES

While on a journey of physical and spiritual healing, a brilliant neurosurgeon is drawn into the world of the mystic arts.

Un ejemplo trabajado con código

Clasificación Multi-etiqueta

Texto	Action	Adventure	Fantasy	Romance
While on a journey of physical and spiritual healing, a brilliant...	1	1	1	0
...

Un ejemplo trabajado con código

Más ejemplos de clasificación

- Clasificación de imágenes (identificar objetos en fotos).
- Diagnóstico médico (clasificar si un tumor es benigno o maligno). Esto puede ser por medio de imágenes, mediciones, etc.
- Reconocimiento de voz (identificar palabras habladas).
- Detección de fraude (identificar transacciones fraudulentas).
- Detección de tópicos (Identificar el tópico de un documento escrito).
- Análisis de sentimientos (Identificar el sentimiento detrás de un texto).

La tarea de clasificación: Planteamiento matemático

Datos de entrada en la clasificación binaria:

$$X = \underbrace{\{x_1, \dots, x_n\}}_{\text{Datos de entrada}} \subset \mathbb{R}^D, \quad Y = \underbrace{\{y_1, \dots, y_n\}}_{\text{Etiqueta de cada dato}}$$

donde $y_j \in \{0, 1\}$.

La tarea de clasificación: Planteamiento matemático

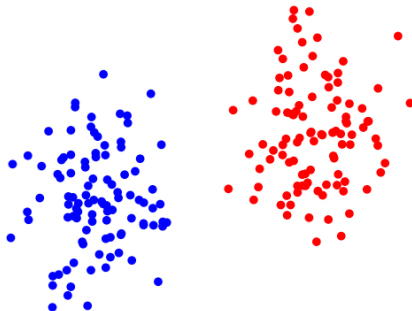
Datos de entrada en la clasificación binaria:

$$X = \underbrace{\{x_1, \dots, x_n\}}_{\text{Datos de entrada}} \subset \mathbb{R}^D, \quad Y = \underbrace{\{y_1, \dots, y_n\}}_{\text{Etiqueta de cada dato}}$$

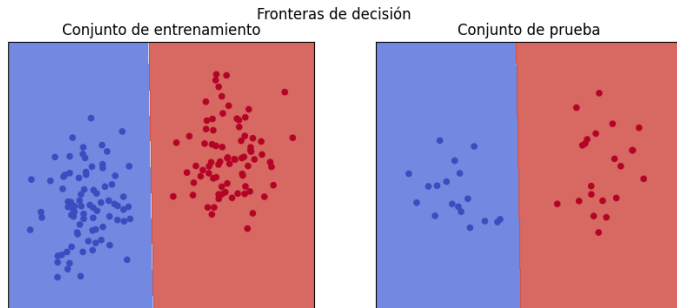
donde $y_j \in \{0, 1\}$.

- **Enfoque ML:** Un clasificador asigna etiquetas a cada dato de entrada.
- **Enfoque geométrico:** Es decir, separa los datos de entrada X en regiones de decisión cuyos límites se llaman fronteras de decisión.

La geometría



La geometría



El algoritmo buscará encontrar la frontera de decisión de acuerdo a diferentes criterios.

Algoritmos

Hay varios métodos:

- SVM (Support Vector Machine)
- Regresión Logística
- Árboles de decisión y árboles aleatorios.
- Naive-Bayes
- K-nearest neighbors.
- Perceptron (Módulo siguiente)
- Redes Neuronales (Módulo siguiente)

¿Cómo lidiar con la clasificación multiclase?

Algunos algoritmos no soportan la clasificación multiclase, sólo la binaria. En estos casos hay dos estrategias para convertir una clasificación multiclase (con k clases diferentes) en varios problemas de clasificaciones binarias:

- **One vs all** (OVA) o **one vs rest** (OVR). Se divide una clasificación multiclase en un problema de clasificación binaria por cada clase. En cada clasificación binaria se analiza si la entidad pertenece a la clase j -sima o no.

¿Cómo lidiar con la clasificación multiclase?

Algunos algoritmos no soportan la clasificación multiclase, sólo la binaria. En estos casos hay dos estrategias para convertir una clasificación multiclase (con k clases diferentes) en varios problemas de clasificaciones binarias:

- **One vs all** (OVA) o **one vs rest** (OVR). Se divide una clasificación multiclase en un problema de clasificación binaria por cada clase. En cada clasificación binaria se analiza si la entidad pertenece a la clase j -sima o no.
- **One vs one** (OVO). Se divide una clasificación multiclase en un problema de clasificación binaria por cada par de clases. En cada clasificación binaria se analiza si la entidad pertenece a la clase i -sima o a la clase j -sima.

¿Cómo lidiar con la clasificación multiclase?

Supongamos que tenemos un conjunto de datos en el que cada instancia puede ser de clase rojo, verde o azul.

- **One vs rest (OVR).**

- Clasificación binaria 1: Rojo, (azul, verde).
- Clasificación binaria 2: Azul, (rojo, verde).
- Clasificación binaria 3: Verde, (azul, rojo).

objeto	color
objeto 1	rojo
objeto 2	verde

¿Cómo lidiar con la clasificación multiclase?

Supongamos que tenemos un conjunto de datos en el que cada instancia puede ser de clase rojo, verde o azul.

- **One vs rest (OVR).**

- Clasificación binaria 1: Rojo, (azul, verde).
- Clasificación binaria 2: Azul, (rojo, verde).
- Clasificación binaria 3: Verde, (azul, rojo).

objeto	rojo	verde	azul
objeto 1	1	0	0
objeto 2	0	1	0

¿Cómo lidiar con la clasificación multiclase?

Supongamos que tenemos un conjunto de datos en el que cada instancia puede ser de clase rojo, verde o azul.

- **One vs one (OVO).**

- Clasificación binaria 1: Rojo, azul.
- Clasificación binaria 2: Rojo, verde.
- Clasificación binaria 3: Azul, verde.

objeto	color
objeto 1	rojo
objeto 2	verde

Un ejemplo ilustrativo...

¿Cómo lidiar con la clasificación multiclase?

Supongamos que tenemos un conjunto de datos en el que cada instancia puede ser de clase rojo, verde o azul.

- **One vs one (OVO).**

- Clasificación binaria 1: Rojo, azul.
- Clasificación binaria 2: Rojo, verde.
- Clasificación binaria 3: Azul, verde.

objeto	rojo/azul	rojo/verde	azul/verde
objeto 1	1	1	-
objeto 2	-	0	0

Un ejemplo ilustrativo...

Table of Contents

- 1 Introducción: La tarea de clasificación
- 2 Métricas de desempeño
- 3 Comparación de algoritmos

Matriz de Confusión Binaria

		Predicted condition	
		Positive (PP)	Negative (PN)
Actual condition	Total population = P + N		
	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Métricas de desempeño

- **Accuracy:** De todos la población, ¿cuántos predije correctamente?

$$A = \frac{TP + TN}{\text{Total}}.$$

- **Recall:** De todos la población positiva, ¿cuántos predije correctamente como positivos?

$$R = \frac{TP}{TP + FN} = TPR.$$

- **Precision:** De todos los que predije como positivos, ¿cuántos son realmente positivos?

$$P = \frac{TP}{TP + FP}.$$

- **F1 score:** Media armónica de la precisión y el recall:

$$2 \frac{P \cdot R}{P + R}$$

Ejemplo

Tenemos la siguiente población $\{+ + - - - -\}$:

- Si nuestro clasificador predice todo como $-$:

real	+	+	-	-	-	-
predicho	-	-	-	-	-	-

Accuracy: 0.66, Recall: 0, Precision: 0.

- Si nuestro clasificador predice todo como $+$:

real	+	+	-	-	-	-
predicho	+	+	+	+	+	+

Accuracy: 0.33, Recall: 1, Precision: 0.33.

Una métrica alta no pinta el panorama completo.

ROC-AUC Score

ROC-AUC Score

La curva paramétrica ROC (Receiver Operating Characteristic) muestra los valores FPR y TPR en varios valores de umbral de probabilidad. El **score AUC** es el area bajo la curva ROC, es una medida de rendimiento para los problemas de clasificación que representa el grado o medida de separabilidad. Indica la capacidad del modelo para distinguir entre clases.

ROC-AUC Score

ROC-AUC Score

La curva paramétrica ROC (Receiver Operating Characteristic) muestra los valores FPR y TPR en varios valores de umbral de probabilidad. El **score AUC** es el area bajo la curva ROC, es una medida de rendimiento para los problemas de clasificación que representa el grado o medida de separabilidad. Indica la capacidad del modelo para distinguir entre clases.

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

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

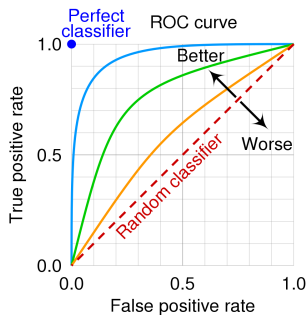
ROC-AUC Score

ROC-AUC Score

La curva paramétrica ROC (Receiver Operating Characteristic) muestra los valores FPR y TPR en varios valores de umbral de probabilidad. El **score AUC** es el area bajo la curva ROC, es una medida de rendimiento para los problemas de clasificación que representa el grado o medida de separabilidad. Indica la capacidad del modelo para distinguir entre clases.

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

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



El valor ROC-AUC es un número $0 \leq s \leq 1$. Entre más grande es s , el clasificador es mejor.

- Si $s = 1$, el clasificador es perfecto.
- Si $s = \frac{1}{2}$, el clasificador es aleatorio.
- Si $s = 0$, el clasificador predice perfectamente las clases *a/ revés*.

Un ejemplo

Umbral: 0.5

y_test	y_pred	probabilidades
0	0	0.048
0	0	0.145
1	1	0.905
0	0	0.24
1	0	0.215
0	0	0.231
0	0	0.116
1	1	0.551
1	0	0.172
1	1	0.803

$$\begin{pmatrix} 5 & 0 \\ 2 & 3 \end{pmatrix}, \quad TPR = 0.6, \quad FPR = 0$$

Un ejemplo

Umbral: 0.2

y_test	y_pred	probabilidades
0	0	0.048
0	0	0.145
1	1	0.905
0	1	0.24
1	1	0.215
0	1	0.231
0	0	0.116
1	1	0.551
1	0	0.172
1	1	0.803

$$\begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix}, \quad TPR = 0.8, \quad FPR = 0.4$$

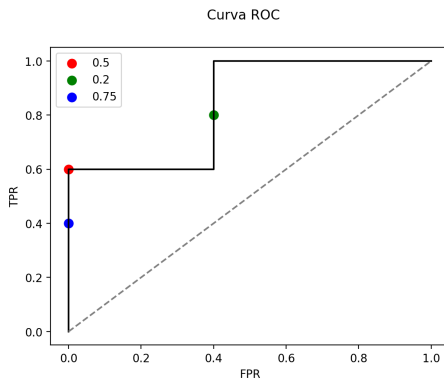
Un ejemplo

Umbral: 0.75

y_test	y_pred	probabilidades
0	0	0.048
0	0	0.145
1	1	0.905
0	0	0.24
1	0	0.215
0	0	0.231
0	0	0.116
1	0	0.551
1	0	0.172
1	1	0.803

$$\begin{pmatrix} 5 & 0 \\ 3 & 2 \end{pmatrix}, \quad TPR = 0.4, \quad FPR = 0$$

Un ejemplo



El área bajo la curva es 0.84.

Table of Contents

- 1 Introducción: La tarea de clasificación
- 2 Métricas de desempeño
- 3 Comparación de algoritmos**

Comparación de algoritmos

Algoritmo	Ventajas	Desventajas
SVM	Útil en alta dimensión ($N < D$)	Datasets grandes
	Flexibilidad	Sensibilidad a outliers
Árboles de Decisión	Capturar relaciones no lineales complejas entre features y target	Sobreajuste
	Interpretabilidad e importancia de features	Sensibilidad a perturbaciones
	Features categóricas	

Comparación de algoritmos

Algoritmo	Ventajas	Desventajas
Random Forest	Robustez y rendimiento	Perdemos interpretabilidad
	Puede manejar datos faltantes	Alto costo computacional
	importancia de features	
Regresión Logística	Interpretabilidad (coeficientes)	Le afectan las features colineales
	Baseline	Puede no funcionar bien con datos desequilibrados
	Puede descubrir relaciones no lineales	Suele no tener buen rendimiento como otros algoritmos