Procesamiento de Lenguaje Natural

Clasificación

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora

Section 1

Introducción

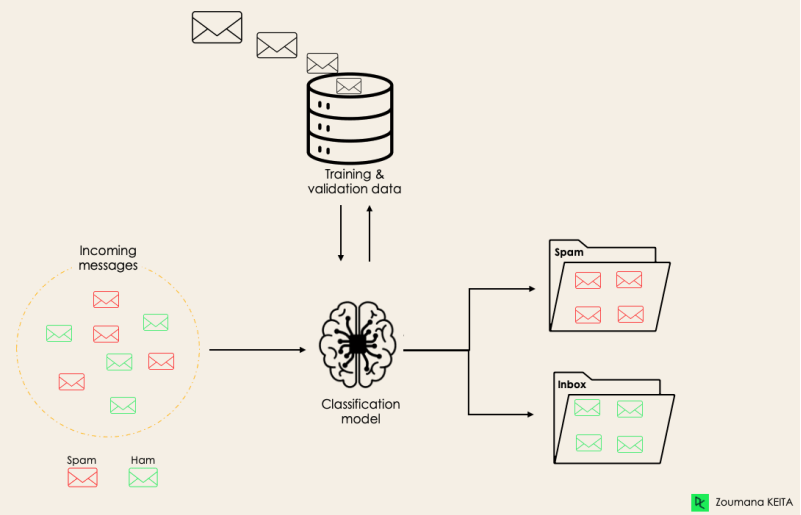
Referencias

- **Chapter I.2, I.4**. Eisenstein, J. (2018). Natural language processing. Jacob Eisenstein.
- **Chapter 4**. Jurafsky, D., Martin, J. H. (2019). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Clasificación

Clasificación

La **clasificación** consiste en asignar categorias a objetos. En el ML supervisado un modelo intenta predecir la etiqueta correcta de datos de entrada.

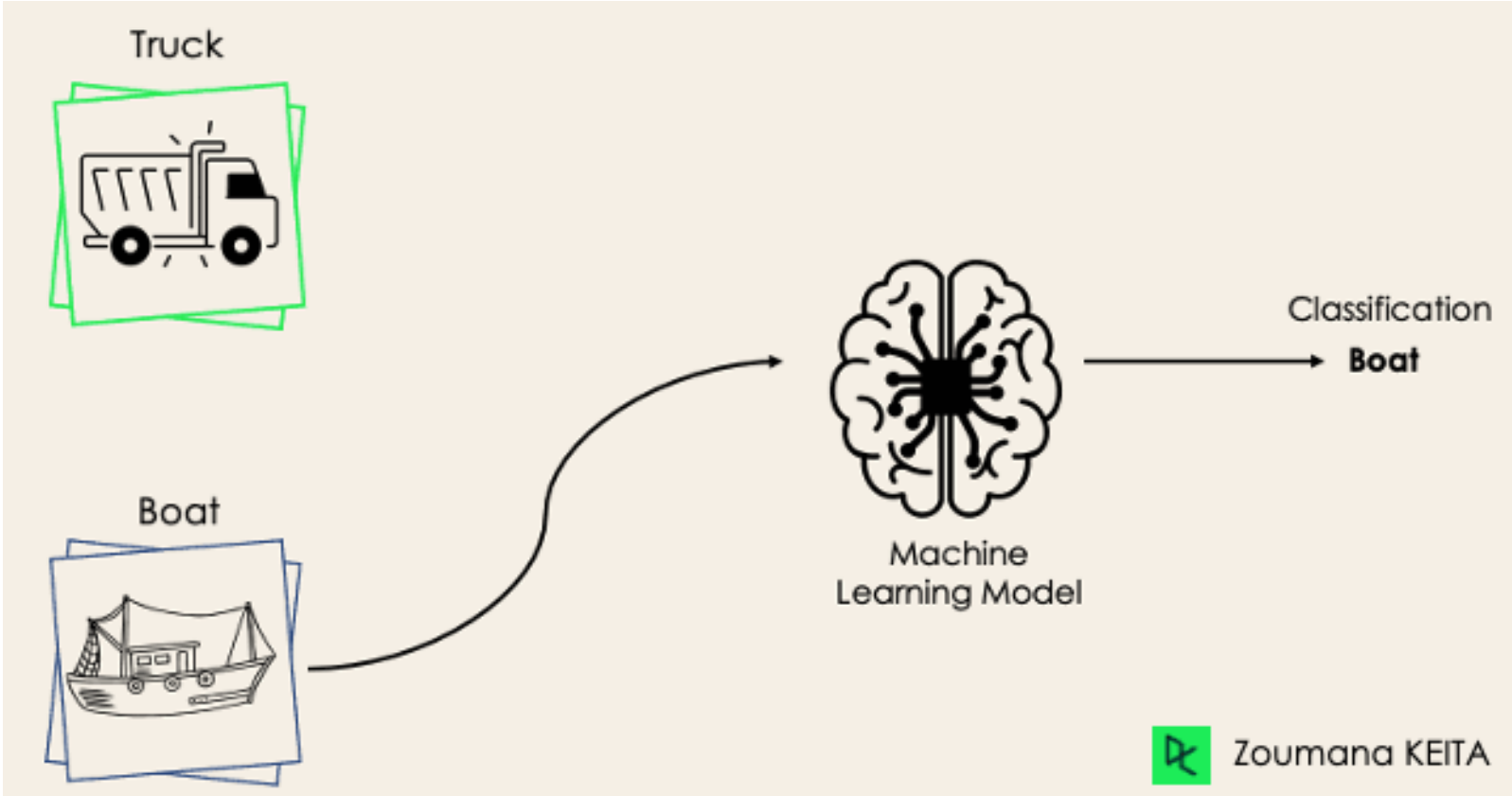


Source

Tipos de Clasificación

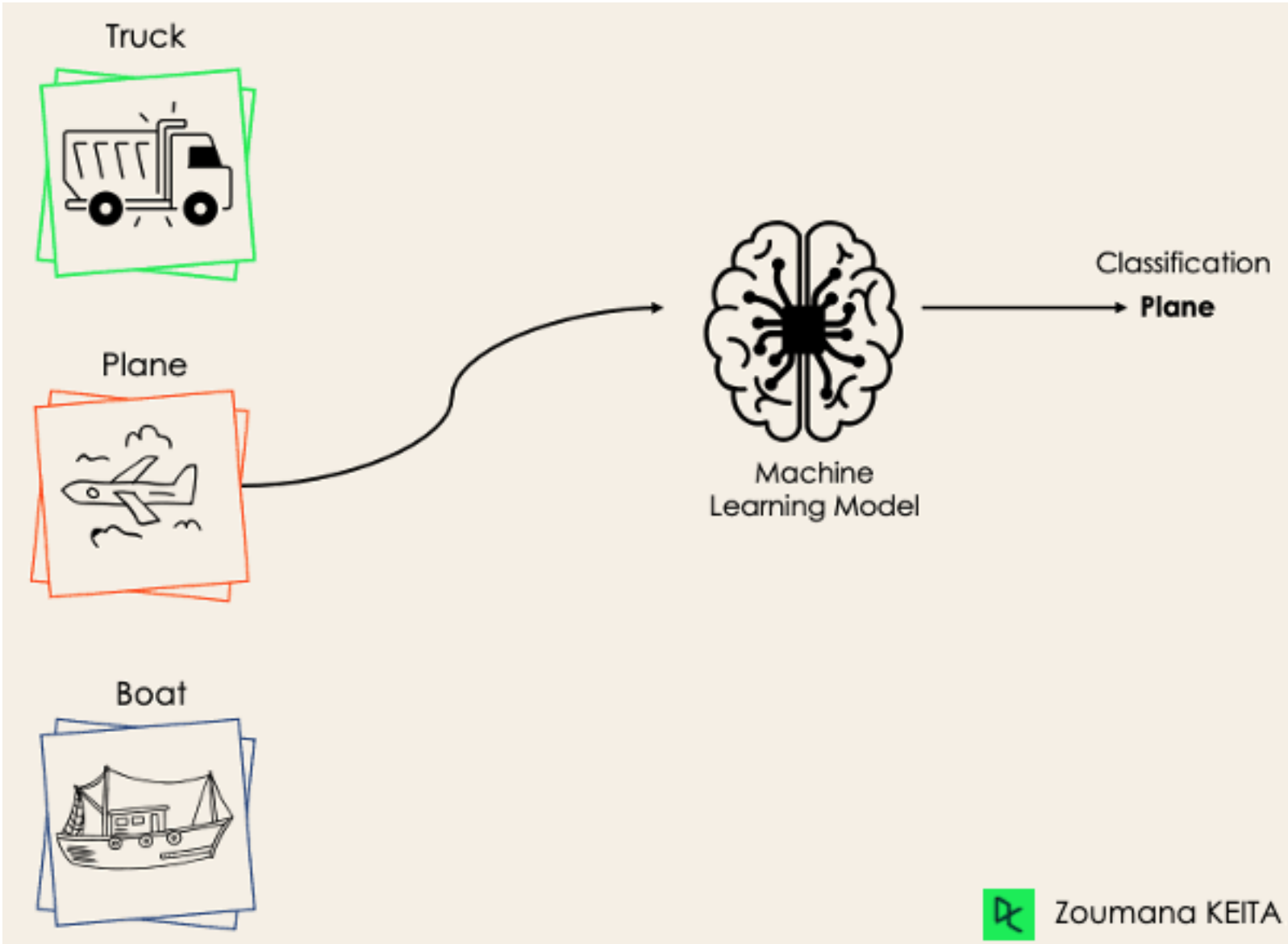
- **Clasificación Binaria**. Cada dato tiene sólo una de dos posibles etiquetas.
- **Clasificación MultiClase**. Cada dato tiene sólo una de varias posibles etiquetas.
- **Clasificación MultiEtiqueta**. Cada dato tiene una o más de varias posibles etiquetas.

Tipos de Clasificación: ejemplos

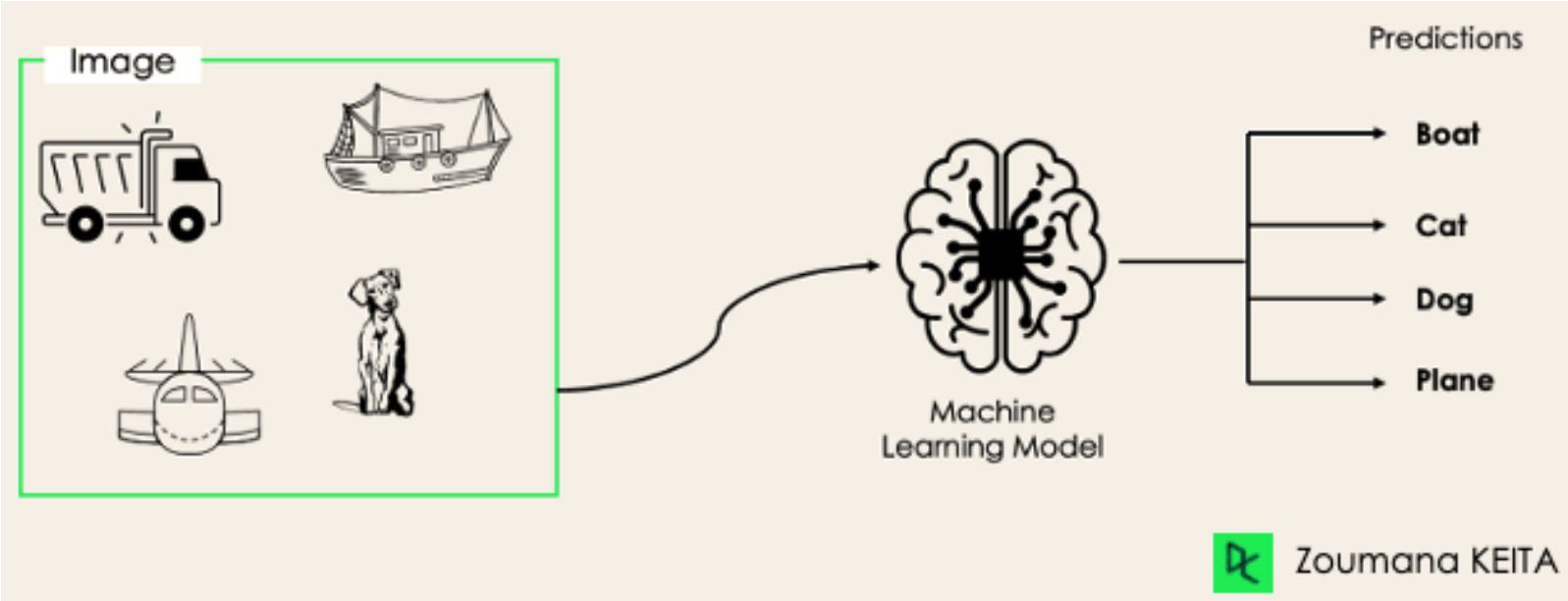


Source

Tipos de Clasificación: ejemplos



Tipos de Clasificación: ejemplos



Source

Ejemplos de clasificación

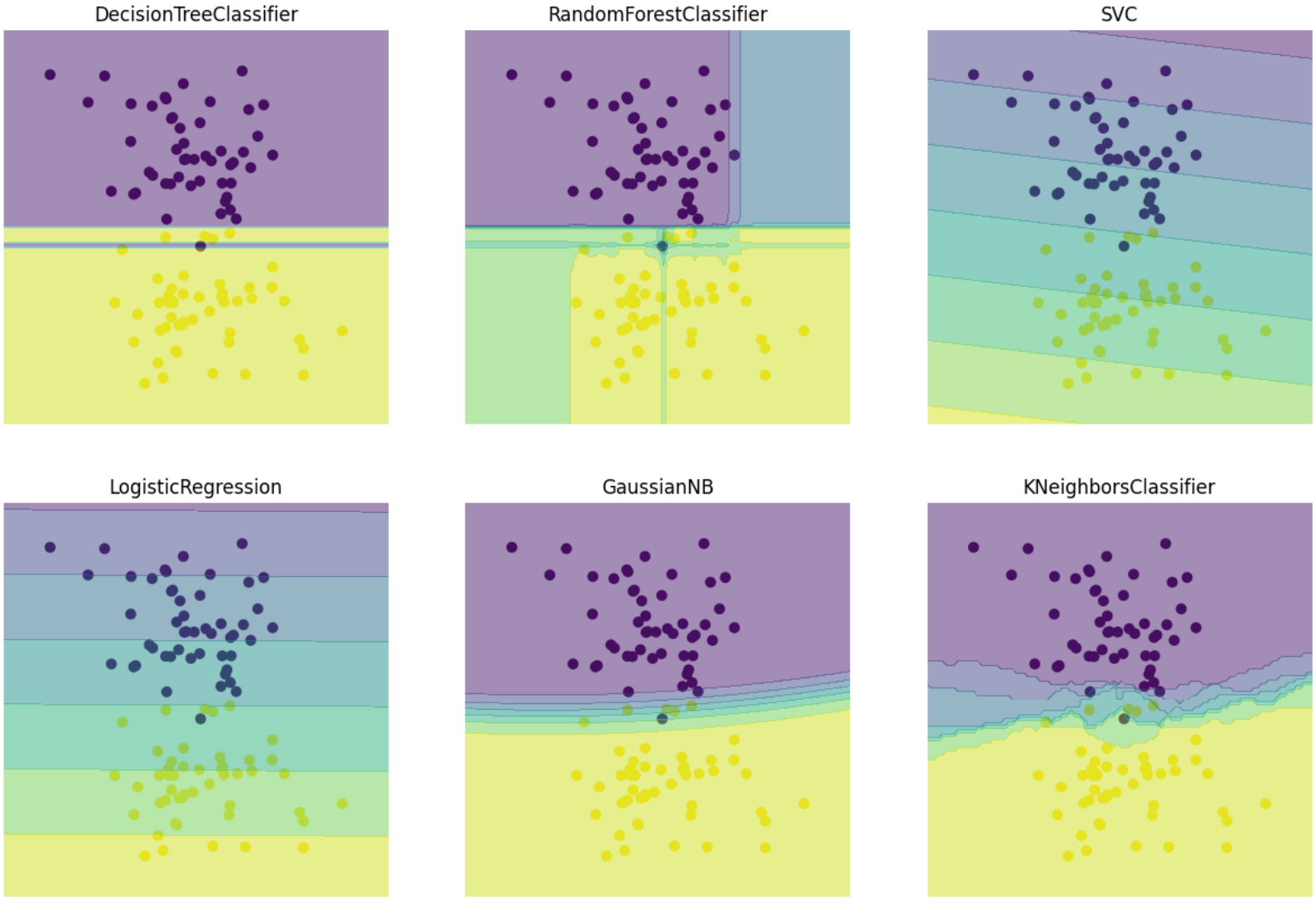
- Clasificación de imágenes (identificar objetos en fotos).
- Diagnóstico médico (clasificar si un tumor es benigno o maligno). Esto puede ser por medio de imágenes, mediciones, etc.
- Reconocimiento de voz (identificar palabras habladas).
- Detección de fraude (identificar transacciones fraudulentas).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

¿Cómo hacemos la clasificación?

- **Métodos basados en reglas**. Las reglas suelen generarse por expertos en el dominio.
- **Métodos de aprendizaje automático**. La lógica para inferir la clase de un objeto se aprende a partir de ejemplos etiquetados correctos previamente.

Algoritmos de Clasificación

- Clasificación Lineal
  - SVM
  - Regresión Logística
  - Perceptrón
  - Clasificador de Mínimos Cuadrados
- Clasificación No lineal
  - Árboles de Decisión
  - Clasificadores de Ensamble
  - K-NN
  - Redes Neuronales
  - **Naive Bayes**


DecisionTreeClassifier
RandomForestClassifier
SVC
LogisticRegression
GaussianNB
KNeighborsClassifier

Evaluación de la tarea de clasificación

¿Cómo evaluamos que tan *bueno* es el desempeño de nuesto modelo de clasificación?

Matriz de Confusión Binaria

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**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: Una sola métrica no es suficiente

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.

Un accuracy alto no pinta el panorama completo.

El precision y recall, a diferencia del accuracy, hacen hincapié en los verdaderos positivos: encontrar las cosas que se supone que estamos buscando.

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.

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

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

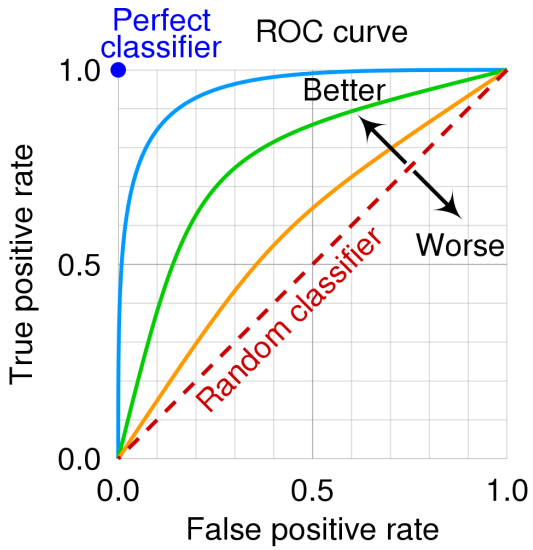
ROC-AUC Score

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 calsificador predice perfectamente las clases *al 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

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

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

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

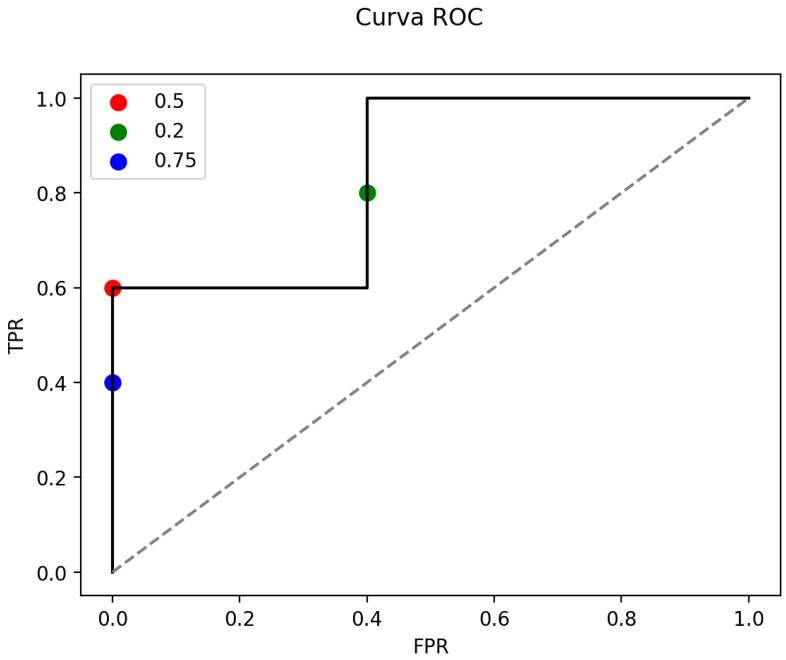
Un ejemplo

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

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

Un ejemplo



El área bajo la curva es 0.84.

Section 2

Clasificación en el PLN

Clasificación en PLN

Consideremos la tarea de clasificación en PLN, es decir, asignar categorías a textos.

- **Análisis de sentimientos**: La extracción del sentimiento, es decir, la orientación positiva o negativa que el escritor expresa hacia algún objeto.

Reseña de una película, un libro o un producto	$\leftrightarrow$	El sentimiento del autor hacia el producto
Editorial o un texto político	$\leftrightarrow$	El sentimiento hacia un candidato o una acción política.
Textos en redes sociales	$\leftrightarrow$	El estado de animo.

Clasificación en PLN

- Detección de SPAM
- Identificación de idioma
- Atribución de autoría
- Detección de tópicos o temática
- Predicción de la siguiente palabra

¿Cómo hacemos la clasificación?

- **Métodos basados en reglas**. Las reglas suelen generarse por expertos en el dominio y pueden incluir coincidencias de palabras o secuencias específicas, patrones sintácticos, patrones léxicos (longitud de palabras, frecuencias).
  - \- Las reglas suelen ser frágiles.
  - \- Las reglas pueden cambiar con el tiempo.
  - \+ Eficientes y no requieren muchos datos.
  - \+ Transparentes y explicables.
- **Métodos de aprendizaje automático**.
- **Métodos Híbridos**: Lexicon-based, ...

Lexicon-based Sentiment Analysis

Algunas palabras tienen una polaridad globalmente reconocida, por ejemplo *bueno*, *malo*, *perfecto*, *feo*. La simple presencia de una de estas palabras en un texto puede ser una pista importante sobre el sentimiento expresado.

"A good tool that works perfectly"

"I had an horrible experience"

En la clasificación Lexicon-based[1], se crea una lista de palabras para cada etiqueta y se clasifica cada documento en función de cuántas palabras de cada lista están presentes.

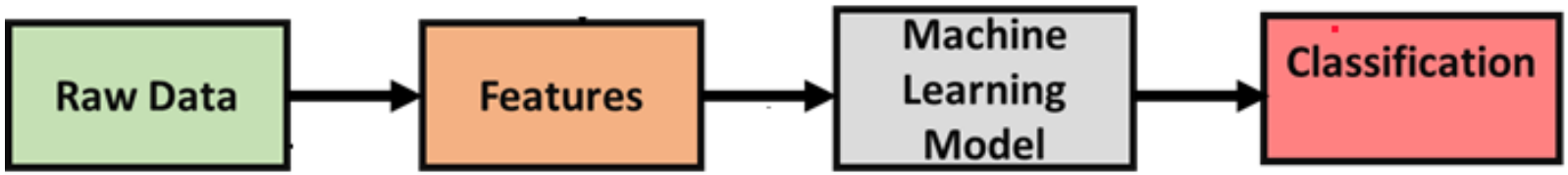
[1] Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. Computational linguistics 37(2), 267–307.

Diccionarios

Existen diccionarios ya recopilados que asocian sentimientos a palabras:

- MPQA
- The General Inquirer lexicon
- SentiWordNet
- Appraisal lexicon

Clasificación en el Machine Learning



¿Quienes son las *features* en el NLP?

- Métodos clásicos: BOW, TF-IDF, ...
- Embeddings: Redes Neuronales, LLMs.

Retomando el clasificador vigente

- Features
- Raw Data
- Evaluación
- Rule-based / Lexicon-based / Machine Learning