Procesamiento de Lenguaje Natural

Machine Learning

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora

Section 1

Introducción

¿Qué es el Machine Learning?

Machine Learning

El **Machine Learning** (Aprendizaje Automático) es un campo de la inteligencia artificial que se enfoca en desarrollar algoritmos que permiten a las computadoras aprender de los datos y mejorar su rendimiento en una tarea específica sin ser programadas explícitamente.

Machine Learning vs Algoritmos tradicionales

Un **algoritmo tradicional** toma una entrada y una lógica en forma de código y genera una salida para resolver un problema.

Machine Learning vs Algoritmos tradicionales

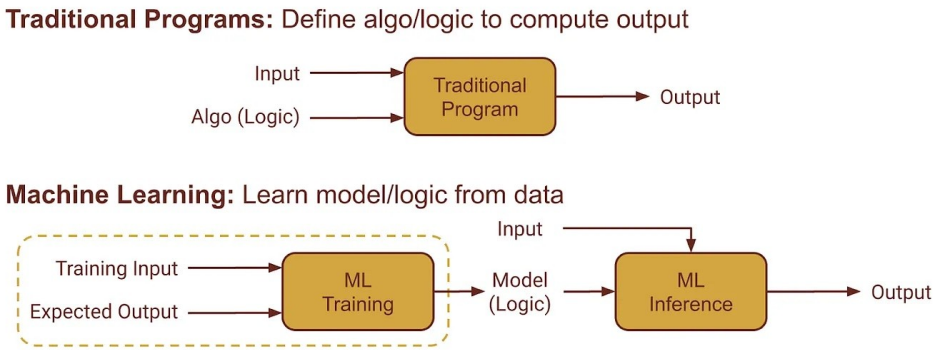
Un **algoritmo tradicional** toma una entrada y una lógica en forma de código y genera una salida para resolver un problema.

Por el contrario, un **algoritmo de aprendizaje automático** toma una entrada y una salida y aprende una lógica que puede utilizarse para trabajar con nuevas entradas y obtener una salida. Esta lógica se obtiene a partir de los patrones presentes en los datos.

Machine Learning vs Algoritmos tradicionales

Un **algoritmo tradicional** toma una entrada y una lógica en forma de código y genera una salida para resolver un problema.

Por el contrario, un **algoritmo de aprendizaje automático** toma una entrada y una salida y aprende una lógica que puede utilizarse para trabajar con nuevas entradas y obtener una salida. Esta lógica se obtiene a partir de los patrones presentes en los datos.



https://www.linkedin.com/pulse/machine-learning-vs-traditional-software-development-ml4devs-gupta

Fases de un programa de Machine Learning

Los programas de aprendizaje automático tienen dos fases distintas:

1. **Entrenamiento**: Las entradas y la salida esperada se utilizan para entrenar y probar varios modelos. Se selecciona el modelo más adecuado. *Entrenar* quiere decir determinar los parámetros adecuados del modelo para producir la salida esperada, a partir de las entradas.
2. **Inferencia o predicción**: El modelo se aplica a nuevos datos de entrada para predecir nuevas salidas, las cuales pueden compararse con las salidas reales.

¿Por qué necesitamos el Machine Learning?

En el enfoque clásico, antes del ML, se usaban algoritmos que procesaban los datos con base en reglas lógicas (`if`, `else`, ...).

¿Por qué necesitamos el Machine Learning?

En el enfoque clásico, antes del ML, se usaban algoritmos que procesaban los datos con base en reglas lógicas (`if`, `else`, ...).

- La lógica para tomar las decisiones es específica de acuerdo al dominio y a la tarea. Pequeños cambios en la tarea requieren rediseñar el sistema.
- El diseño de las reglas requiere un entendimiento profundo del dominio por parte de un experto. Estas reglas pueden ser muy complicadas.

¿Por qué necesitamos el Machine Learning?

En el enfoque clásico, antes del ML, se usaban algoritmos que procesaban los datos con base en reglas lógicas (`if`, `else`, ...).

- La lógica para tomar las decisiones es específica de acuerdo al dominio y a la tarea. Pequeños cambios en la tarea requieren rediseñar el sistema.
- El diseño de las reglas requiere un entendimiento profundo del dominio por parte de un experto. Estas reglas pueden ser muy complicadas.

Ejemplos:

ML	Tradicional
Detectar correo SPAM	Integrar numéricamente una función
Reconocer rostros en una imagen	

Un ejemplo: Detección de SPAM

Enfoque Tradicional

Enfoque Machine Learning

Un ejemplo: Validar direcciones de correo electrónico

Enfoque Tradicional	Enfoque Machine Learning
Cambio en el tipo de documentos.	

¿Qué tareas puede resolver el Machine Learning?

Algunas de las tareas que pueden resolver los métodos de Machine Learning son

1. Regresión.
2. Clasificación.
3. Clustering (segmentación).
4. Traducción automática.
5. Detección de anomalías.
6. Síntesis y muestreo.
7. Generación (texto, imágenes).

¿Por qué es relevante Machine Learning para NLP?

El lenguaje humano es complejo y ambiguo

- El lenguaje natural tiene múltiples significados, contextos y matices.
- Las reglas explícitas (programación tradicional) no son suficientes para capturar esta complejidad.

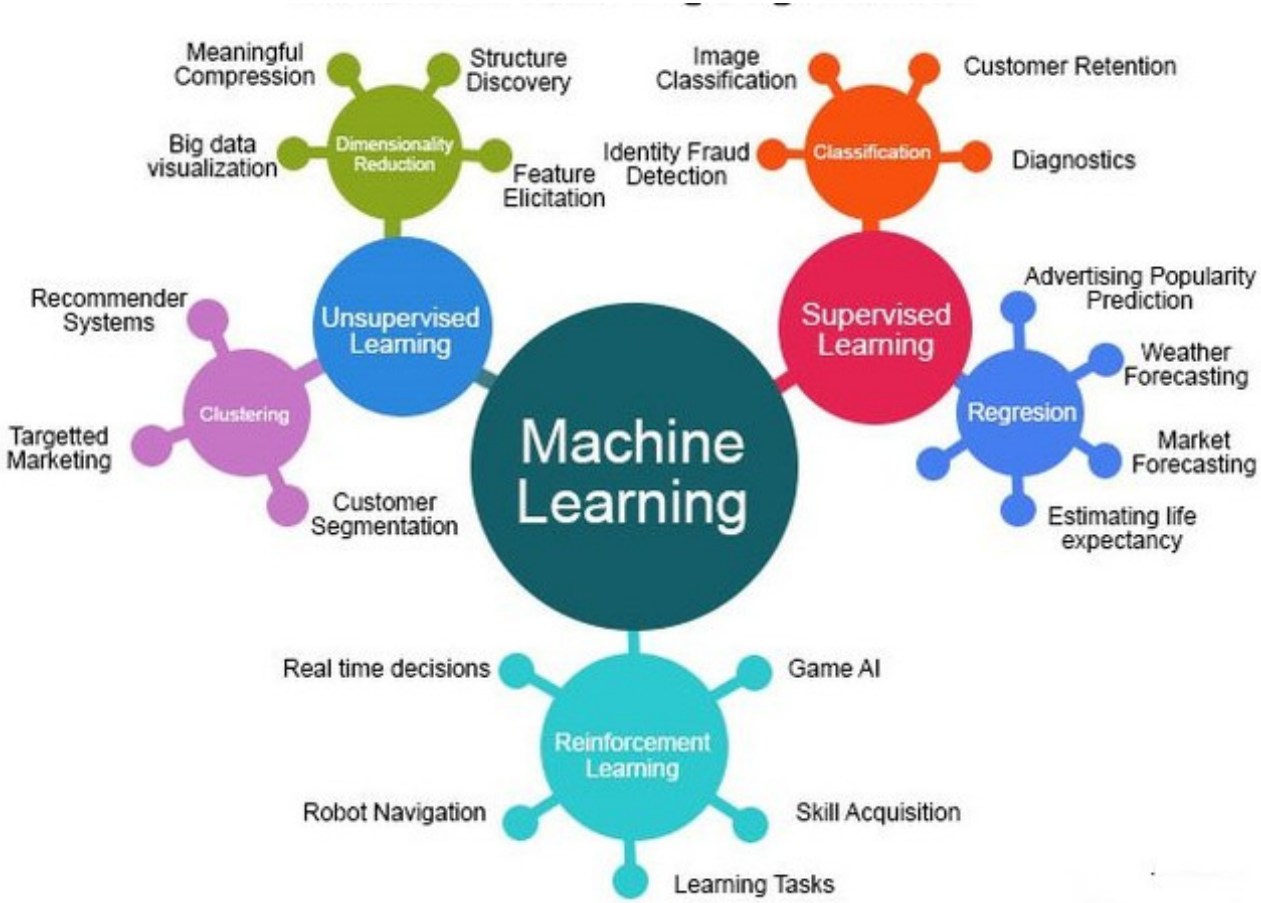
El Machine Learning permite a las computadoras aprender patrones:

- Los algoritmos de ML pueden identificar patrones en grandes cantidades de datos textuales.
- Esto permite a las máquinas tomar decisiones basadas en datos, no en reglas predefinidas.

Los tres paradigmas del Machine Learning

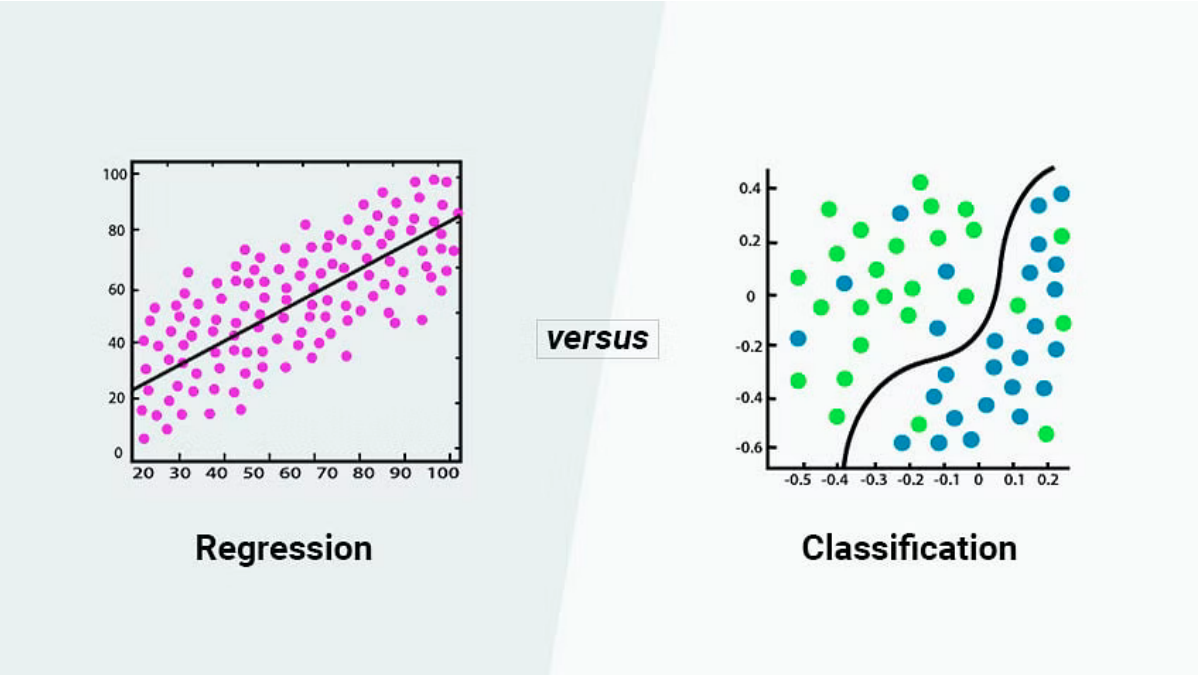
1. **Aprendizaje supervisado (Supervised Learning)**. El modelo aprende de ejemplos etiquetados para hacer predicciones de nuevos datos.
2. **Aprendizaje no supervisado (Unsupervised Learning)**. El modelo encuentra patrones o estructuras intrínsecas en los datos sin etiquetar.
3. **Aprendizaje por refuerzo (Reinforcement Learning)**. El modelo aprende a tomar decisiones en un entorno para maximizar una recompensa acumulativa a través de prueba y error.

Los tres paradigmas del Machine Learning



Aprendizaje Supervisado

1. **Regresión**. La regresión es una tarea en la que se predice un valor numérico continuo a partir de datos de entrada.
2. **Clasificación**. La clasificación es una tarea en la que se predice (o asigna) una clase a cada dato de entrada.

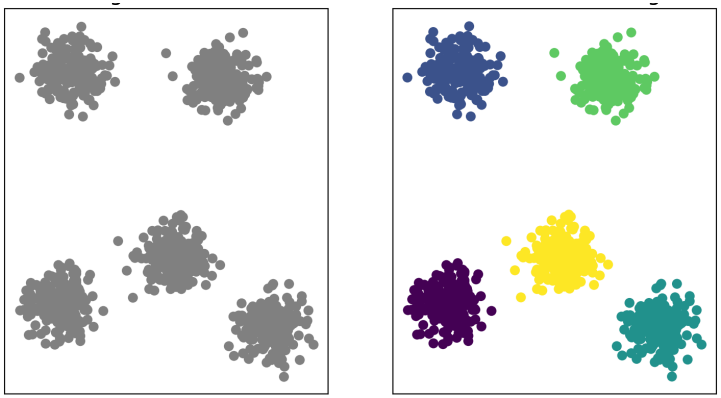


Aprendizaje Supervisado en el NLP

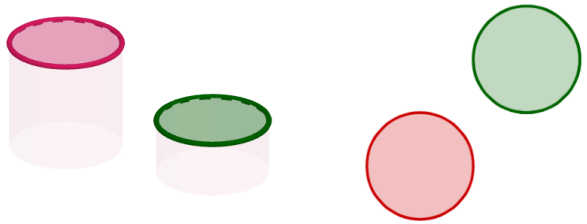
1. Regresión:
   1. Predicción de Puntuación de Sentimientos
   2. Predicción de Popularidad de Contenido
   3. Predicción de Dificultad de Lectura
   4. Predicción de Tiempo de Lectura
   5. Predicción de Calidad de Traducción
   6. Predicción de Relevancia de Documentos
2. Clasificación:
   1. Análisis de Sentimientos
   2. Identificación de SPAM
   3. Identificación del idioma
   4. Identificación de tópicos

Aprendizaje No Supervisado

Clustering. Agrupa datos similares en clusters basándose en sus características. Su objetivo es descubrir patrones inherentes en los datos.



Reducción de dimensionalidad. Transforma datos de alta dimensión a un espacio de menor dimensión, preservando la información más relevante. Su objetivo es simplificar los datos sin perder su esencia.



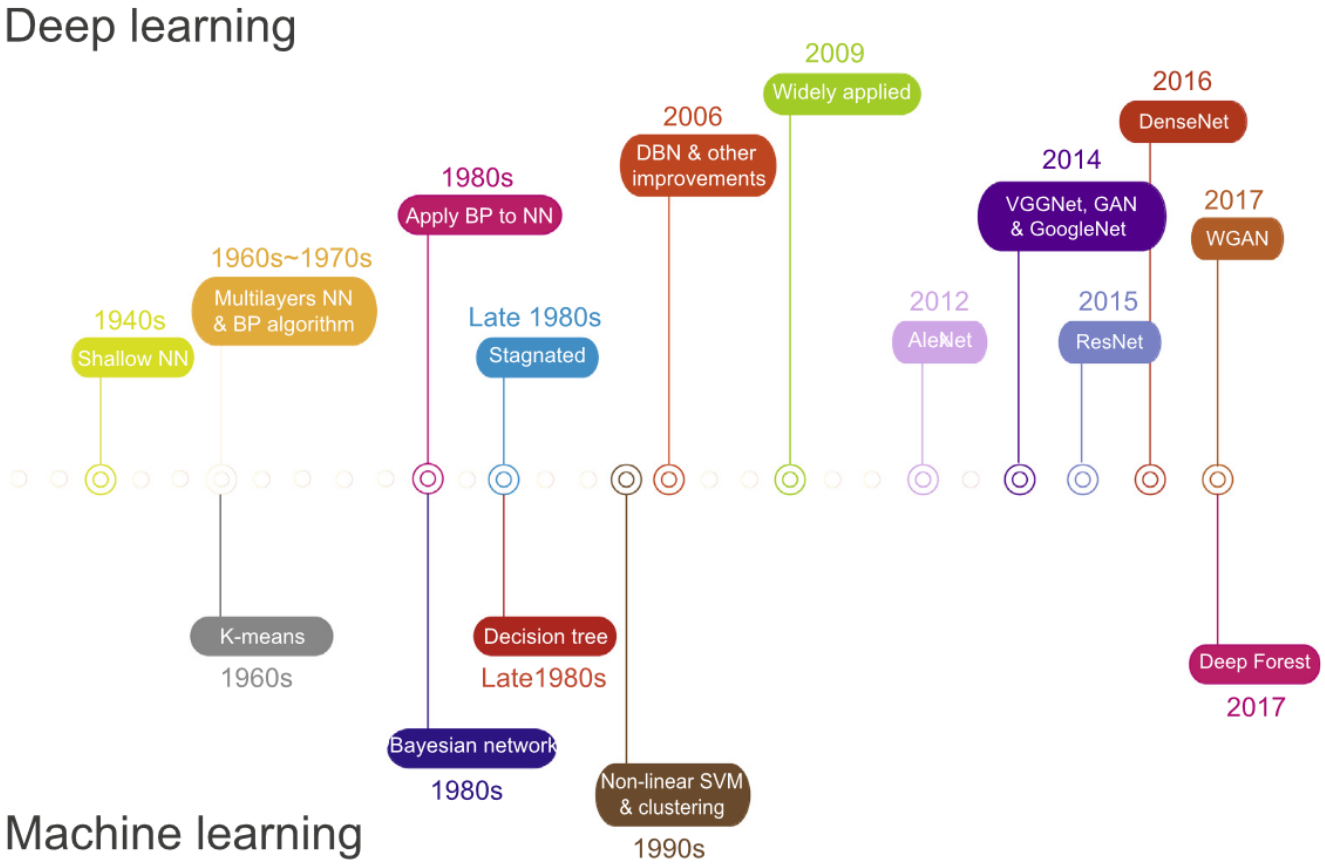
Aprendizaje Supervisado en el NLP

1. Clustering:
   1. Modelado de tópicos
   2. Agrupación de Palabras por Similitud Semántica
   3. Organización de Reseñas por Categorías
   4. Clustering de Noticias por Eventos
   5. Agrupación de Preguntas Frecuentes
2. Reducción de dimensionalidad:
   1. Visualización de Textos en 2D o 3D
   2. Compresión de Embeddings de Palabras
   3. Mejora de la Eficiencia en Modelos de NLP
   4. Reducción de Ruido en Representaciones de Texto
   5. Integración de Datos Multimodales

Aprendizaje por Refuerzo

1. Navegación en vehículos autónomos.
2. Texto predictivo.
3. Sistemas de recomendación.
4. Videojuegos.
5. Chatbots Conversacionales.
6. Traducción Automática.

Timeline de los algoritmos de Machine Learning



https://doi.org/10.1016/j.gpb.2017.07.003

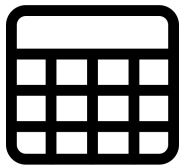
Section 2

Componentes del Machine Learning

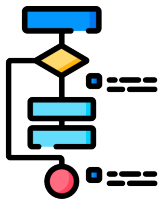
Componentes del Machine Learning

Datos

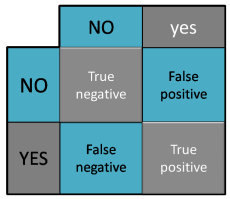
Variables (features)



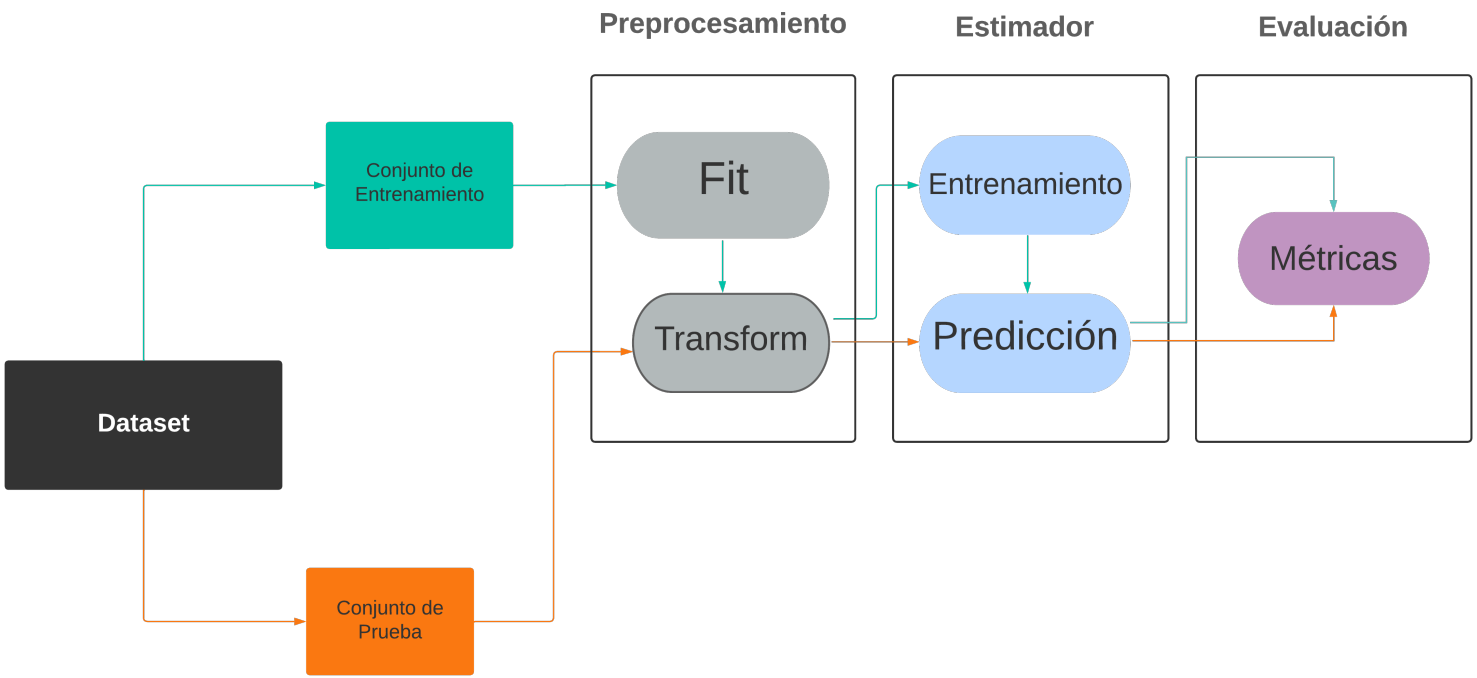
Algoritmos



Métricas



Workflow del Machine Learning



Subsection 1

Datos

Datos

Los datos pueden tener muchas formas diferentes:

- Tablas estructuradas.
- Imágenes.
- **Texto**.
- Archivos de audio.
- Archivos de video.

A un conjunto de datos, se le llama **dataset**. **Kaggle** tiene una colección grande, al igual que **Scikit-Learn**.

Subsection 2

Features y Preprocesamiento

Features

Los datasets suelen organizarse en tablas donde cada fila representa una entidad y cada columna una característica de esa entidad. Algunos tipos comunes de características (variables) son:

- **Numéricas**
- **Categóricas**

Kaggle

No todos los datos se representan naturalmente de esta manera. **El texto no tiene una estructura tabular** inherente. Para poder utilizarlo en modelos de ML, necesitamos transformarlo en características numéricas o categóricas.

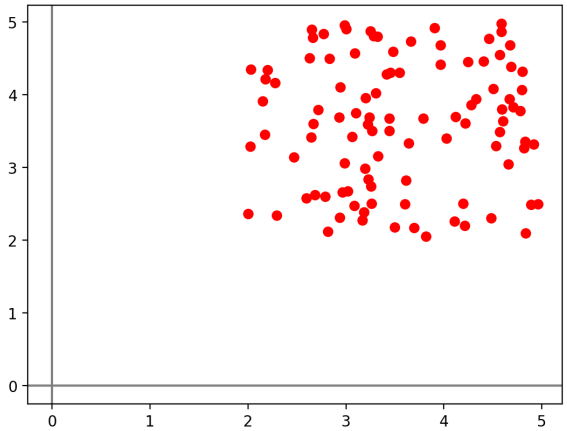
Preprocesamiento

El preprocesamiento es el paso en el que los datos se **transforman** o **codifican** para adaptarlos a un formato que permita a los algoritmos de Machine Learning procesarlos de manera *más eficiente y efectiva*.

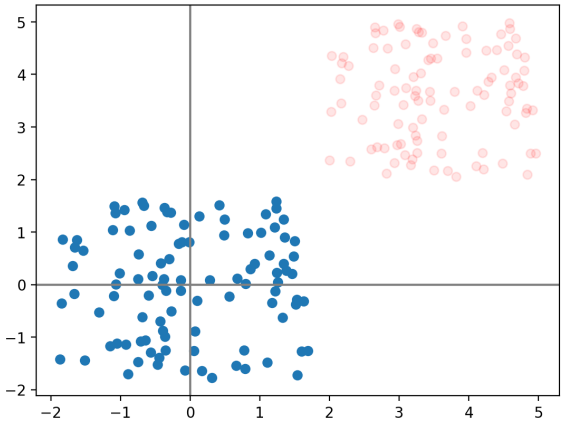
- **Limpieza de datos**.
- Normalización de datos.
- **Transformación de datos**.
- Imputación de datos perdidos.
- Integración de datos.
- Análisis del ruido.

Preprocesamiento

Datos originales



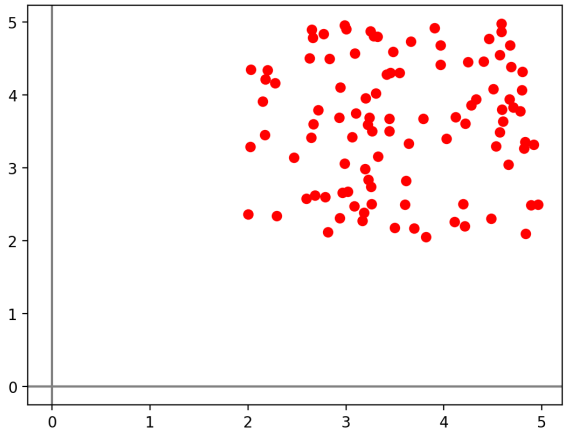
Preprocesamiento



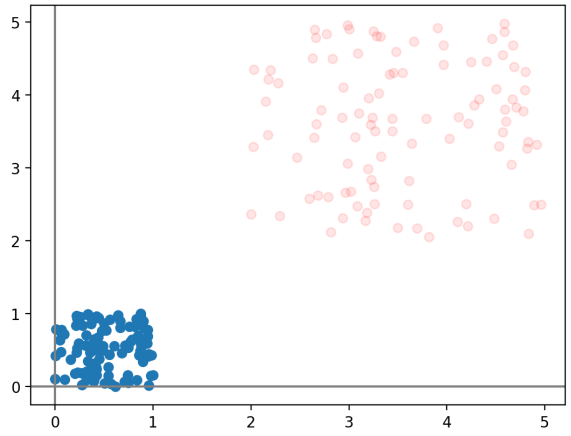
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Preprocesamiento

Datos originales



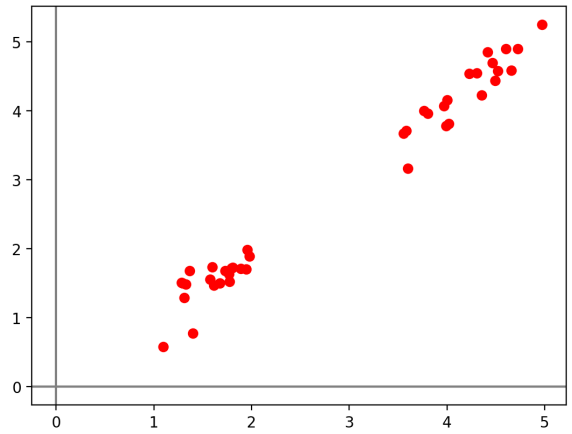
Preprocesamiento



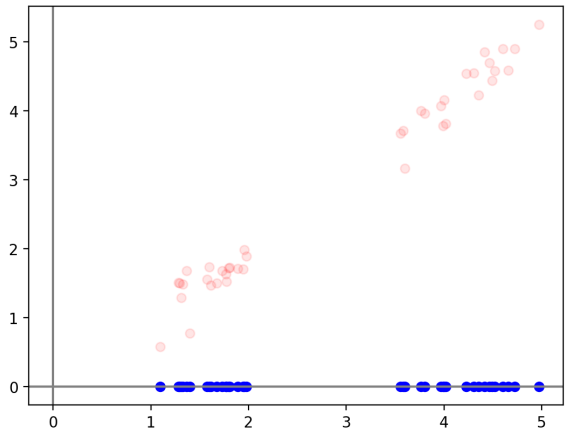
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Preprocesamiento

Datos originales



Preprocesamiento



https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

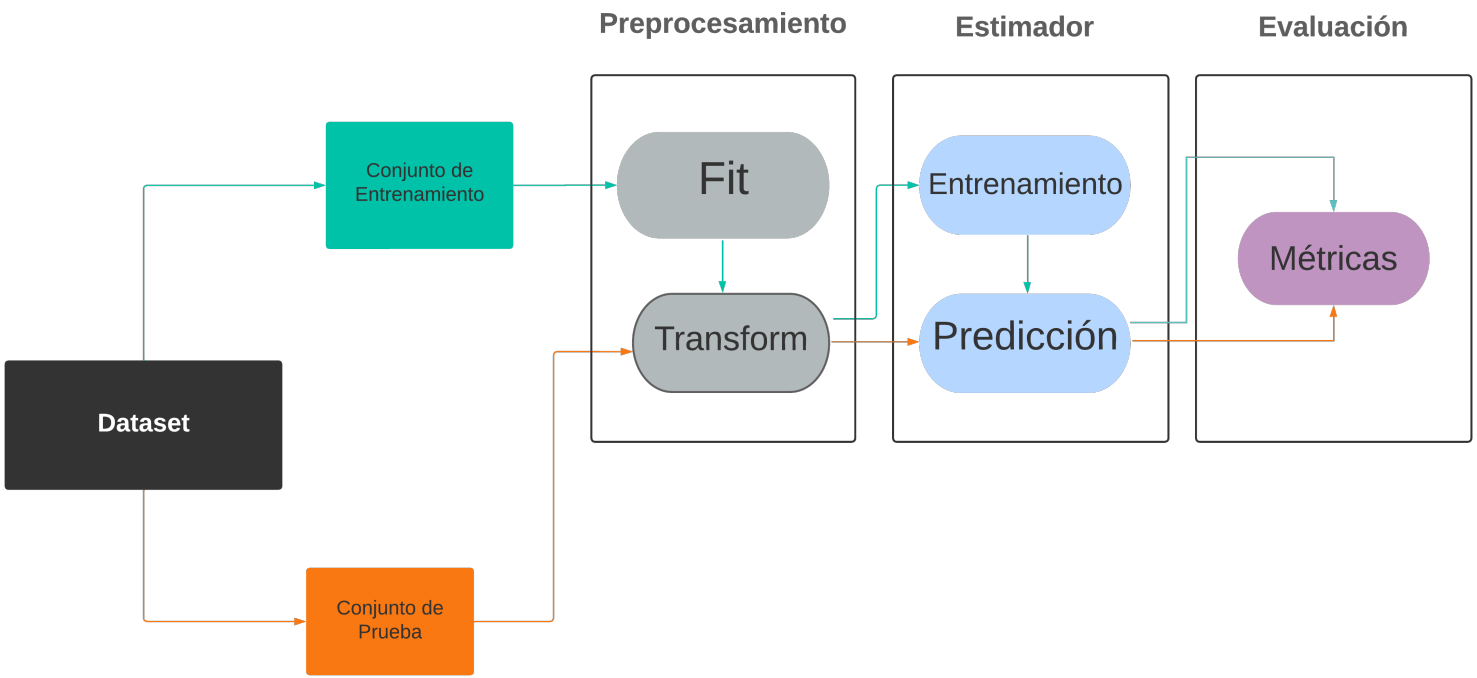
Subsection 3

Algoritmos

Subsection 4

Algoritmos

Workflow del Machine Learning: Algoritmos



Algoritmos

Un **algoritmo de Machine Learning** es la técnica que permite a una computadora aprender a partir de datos y tomar decisiones o hacer predicciones basadas en esa información.

Existen varios tipos de algoritmos de Machine Learning, dependiendo del tipo de tarea que buscan modelar:

- Aprendizaje supervisado.
- Aprendizaje no supervisado.
- Aprendizaje por refuerzo.

Algoritmos


100+ MACHINE LEARNING ALGORITHMS
By @TheInsaneApp
CLUSTERING
- k Means
- k Medians
- BIRCH
- Fuzzy C-Means
- Fuzzy K-Modes
- Mini Batch K-Means
- DBSCAN
- Fuzzy clustering
- Mean-shift
- Optics algorithm
- Expectation Maximization
- Hierarchical Clustering
- Minimum spanning tree
DIMENSIONALITY REDUCTION
- PCA Algorithm
- PCR Algorithm
- PLSR Algorithm
- Sammon Mapping
- MDS Algorithm
- Projection Pursuit
- LDA Algorithm
- ICA Algorithm
- NMF Algorithm
- RDA Algorithm
- MDA Algorithm
- PLSDA Algorithm
- QDA Algorithm
- CCA Algorithm
- Diffusion Map
REGRESSION
- Ordinary Least Squares Regression
- Linear Regression
- Logistic Regression
- Stepwise Regression
- MARS Algorithm
- Locally Estimated Scatterplot Smoothing
DECISION TREE
- Conditional Decision Trees
- Decision Stump
- C4.5 and C5.0
- CHAID
- CART
- ID3
- M5
BAYESIAN
- AODE
- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bayesian Belief Network
- Bayesian Network
REINFORCEMENT LEARNING
- Q Learning
- SARSA Algorithm
- Deep Q-network
- Learning Automata
- DDPG Algorithm
- NAF Algorithm
- A3C Algorithm
- TRPO Algorithm
- PPO Algorithm
- Constructing Skill Trees
REGULARIZATION
- Ridge Regression
- Elastic Net
- LASSO
- LARS
EMSEMBLE
- GBDT
- GBRT
- Boosting
- Bagging
- AdaBoost
- Random Forest
- Blending Algorithm
- Stacked Generalization
- Gradient Boosting Machines
ANNs
- Perceptron
- MLP Algorithm
- Back-Propagation
- Stochastic Gradient Descent
- Hopfield Network
- RBFN Algorithm
OTHER ML ALGORITHMS (Part 1)
- ALOPEX
- CN2 algorithm
- FastICA
- Feature selection Algorithms
- Linde Buzo Gray Algorithm
- Forward Backward algorithm
- Algorithm Accuracy Evaluation
- Performance Measures
- Optimization Algorithm
- Dynamic Time Warping
- Local Outlier Factor
- Logic Learning Machine
- Markov Chain Monte Carlo
- T Distributed Stochastic neighbor Embedding
- Wake Sleep Algorithm
- Prefrontal Cortex Basal Ganglia Working Memory
RULE SYSTEM
- Cubist
- OneR
- ZeroR
- Ripper
OTHER ML ALGORITHMS (Part 2)
- LogitBoost
- Sparse PCA
- Structured kNN
- WMA Algorithm
- GeneRec
- Leabra
- RProp
DEEP LEARNING
- CNN
- RNNs
- LSTMs
- Stacked Auto Encoders
- Deep Boltzmann Machine
- Deep Belief Networks
INSTANCE BASED
- Learning Vector Quantization
- K Nearest Neighbor
- Self-Organizing Map
- Locally Weighted Learning
- Support Vector Machines
ARL
- Apriori Algorithm
- Eclat Algorithm
- FP Growth Algorithm

Subsection 5

Validación

Cross Validation

Validación Cruzada

La validación cruzada es una técnica de validación de modelos para evaluar cómo se generalizarán los resultados de un análisis estadístico a un conjunto de datos independiente. La validación cruzada es un método de remuestreo que utiliza diferentes partes de los datos para probar y entrenar un modelo en diferentes iteraciones.

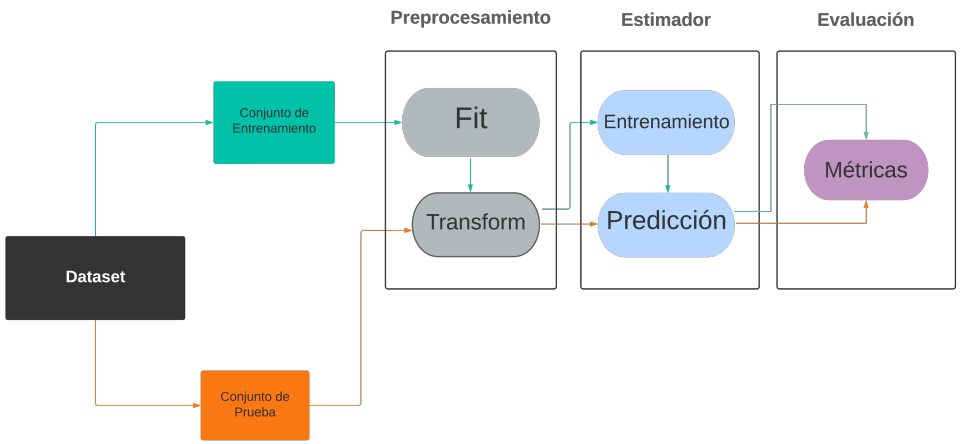
Es necesario tener una validación de la estabilidad de cualquier modelo de Machine Learning. Es decir, ¿qué tan bien podemos esperar que sea su rendimiento en datos que no ha visto?

Tecnicas de validación

- **Validación**: Evaluación del desempeño del modelo en los datos de entrenamiento.

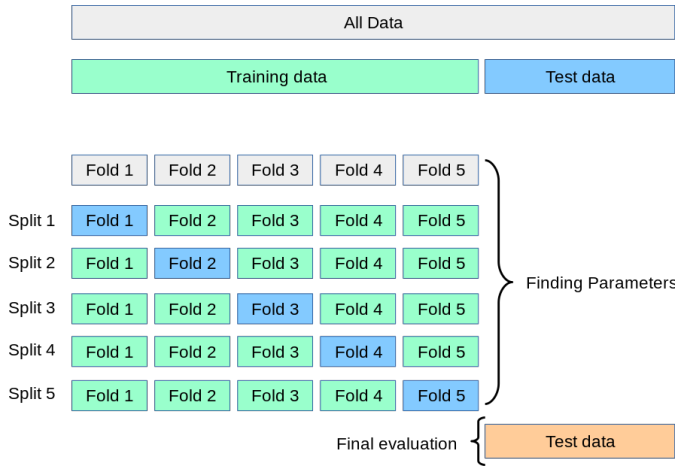
Tecnicas de validación

- **Validación**: Evaluación del desempeño del modelo en los datos de entrenamiento.
- **Conjunto de prueba**: Reservar una parte del conjunto de datos para ser usada como conjunto de prueba.



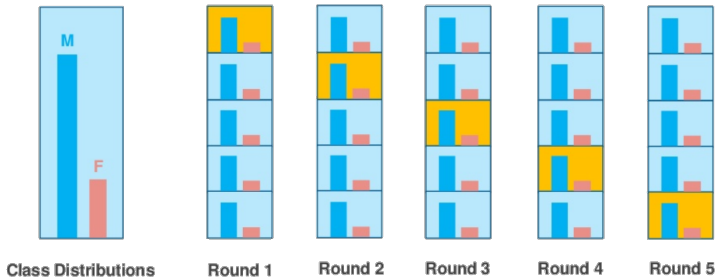
Tecnicas de validación

- **K-Fold Cross Validation**: Los datos se dividen en k subconjuntos, una de las partes se usa como conjunto de prueba y las demás como entrenamiento. Se repite este método k veces, de forma que cada vez, uno de los k subconjuntos se utiliza como conjunto de prueba y los otros $k-1$ subconjuntos, como conjunto de entrenamiento. La estimación del error se promedia sobre las k pruebas para obtener la eficacia total de nuestro modelo.



Tecnicas de validación

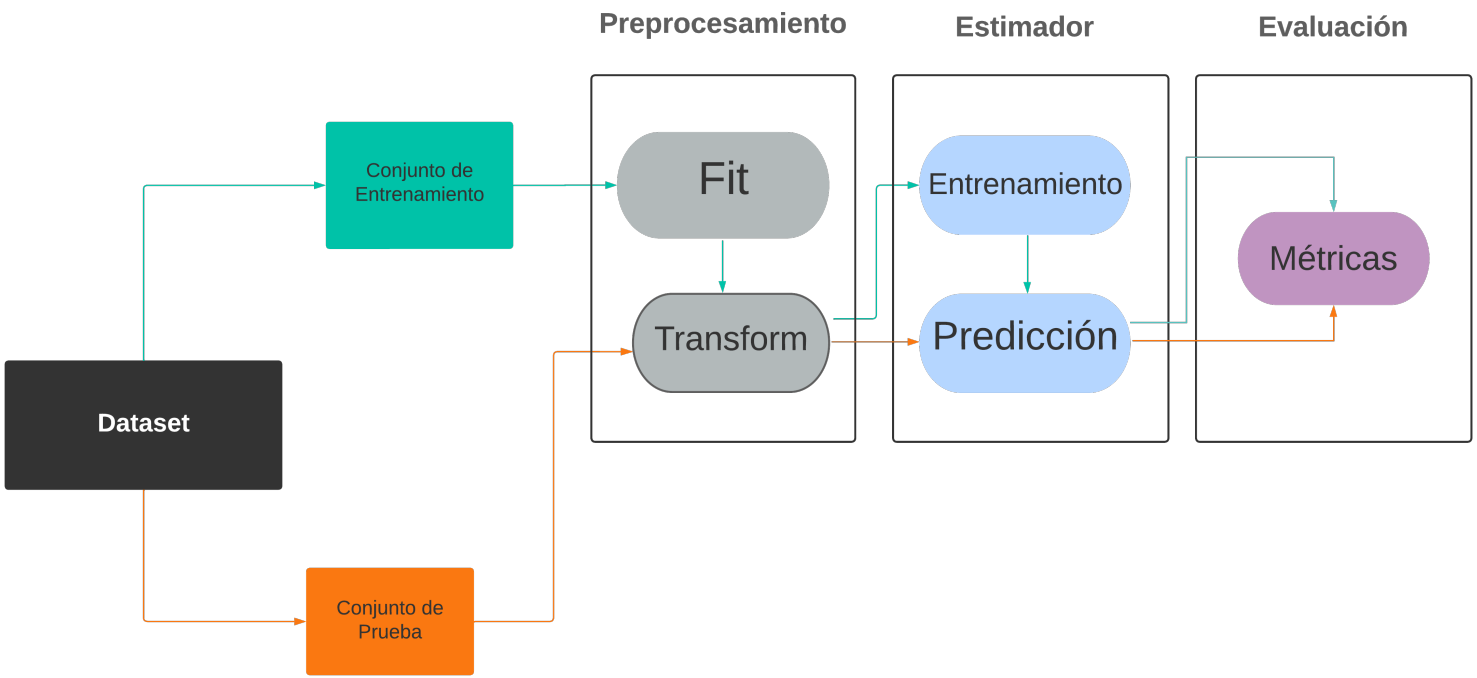
- **Stratified K-Fold Cross Validation**: Variación de la validación cruzada K-fold normal, en lugar de que las divisiones sean completamente aleatorias, la proporción entre las clases objetivo es la misma en cada uno de los k subconjuntos que en el conjunto de datos completo.



Subsection 6

Métricas de Rendimiento

Workflow del Machine Learning: Métricas de rendimiento



Métricas de desempeño: Ejemplo de clasificación

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	?
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.	?
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.	?

Métricas de desempeño: Ejemplo de clasificación

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

Métricas de desempeño

Las **métricas de desempeño** dan cuenta del desempeño del modelo entrenado. Estas funciones varían de acuerdo al tipo de tarea, **suelen ser funciones fácilmente interpretables (porcentajes, conteos, diferencias, etc.)**.

- Regresión: MSE, MAE.
- Clasificación: Accuracy, precision, recall, F1-score, ROC-AUC.
- Clustering: AMI, MI, silhoutte score.
- Algunas propias del NLP: Preplexity, entropy, coherence.
- ...

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


```
Terrible action movie in which lead Franco Nero exchanges his cowboy hat, gun belt and the coffin he
--------------------
In terms of quality movies, this isn't one of them. It's actually the first Chuck Norris movie I've
--------------------
Fans of the Pink Panther, Naked Gun, or Get Smart will certainly enjoy this farce that won one César
--------------------
I am sick of series with young and clueless people, talking about their "problems" all the time, sel
--------------------
Thanks to this film, I now can answer the question, "What is the worst movie you have ever seen?"<br
--------------------
```

Ejemplo

	review	sentiment
4851	Terrible action movie in which lead Franco Ner...	negative
18573	In terms of quality movies, this isn't one of ...	negative
44402	Fans of the Pink Panther, Naked Gun, or Get Sm...	positive
49603	I am sick of series with young and clueless pe...	positive
28873	Thanks to this film, I now can answer the ques...	negative

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.

Funciones de costo

Una **función de pérdida**, o **función de costo**, es una función que asigna un evento o los valores de una o más variables a un número real que representa intuitivamente algún *costo* asociado al evento. Un problema de optimización trata de minimizar una función de pérdida.

Un algoritmo de Machine Learning busca minimizar o maximizar esta función cambiando sus parámetros internos. Frecuentemente se usa el **descenso de gradiente** para este fin, por lo tanto, tipicamente se requiere de una función de costo diferenciable o convexa.

- Regresión: MSE, RMSE, MAE.
- Clasificación: 0-1, binaria asímetrica, entropia cruzada, Hinge loss.

Diferencia entre función de costo y métrica de desempeño

Típicamente son funciones diferentes, bajo ciertas condiciones se puede usar la misma.

- **Usando la función de costo como métrica de desempeño**: puede ser confusa de interpretar.
- **Usando la métrica de desempeño como función de costo**: puede no ser posible si no es diferenciable o convexa.

Resumiendo

Un problema de Machine Learning consiste en los siguientes pasos:

- **Recopilación de datos:** Los datos deben ser suficientes y representativos del problema que se busca resolver.
- **Preprocesamiento:** Limpiar los datos para eliminar ruido, valores faltantes, valores atípicos, y los prepara para su uso en el modelo.
- **Selección del algoritmo**.
- **Entrenamiento del modelo:** Utiliza el conjunto de datos de entrenamiento para entrenar el modelo elegido. La mejora se rige usando la **función de costo**.
- **Evaluación del modelo:** Evalúa el modelo utilizando el conjunto de datos de prueba. Esto se hace con la **métrica de rendimiento**.
- **Implementación, Monitoreo y Mantenimiento.**