

Algoritmos de Aprendizaje Automático

Supervisado y No Supervisado

Con fórmulas, nomenclatura y métricas de evaluación.

Aprendizaje Supervisado

El modelo aprende con **datos etiquetados**.

Algoritmos principales:

- Regresión Lineal / Polinómica
- Regresión con Regularización (Ridge, Lasso)
- Árboles de Decisión
- Random Forest
- Regresión Logística
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Naive Bayes

Regresión Lineal

- **Problemas que resuelve:** Regresión (valores continuos).
- **Fundamento:**

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- **Ventaja:** Muy interpretable.
- **Desventaja:** Solo relaciones lineales.

Regresión Polinómica

- **Problemas que resuelve:** Regresión con relaciones no lineales.
- **Fundamento:**

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n + \varepsilon$$

Nomenclatura:

- n : grado del polinomio.
- Resto igual a la regresión lineal.
- **Ventaja:** Modela curvas complejas.
- **Desventaja:** Riesgo de sobreajuste.

Ridge y Lasso

- Problemas que resuelve: Regresión con regularización.
- Fundamento:

Ridge

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Lasso

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Ridge y Lasso

Nomenclatura:

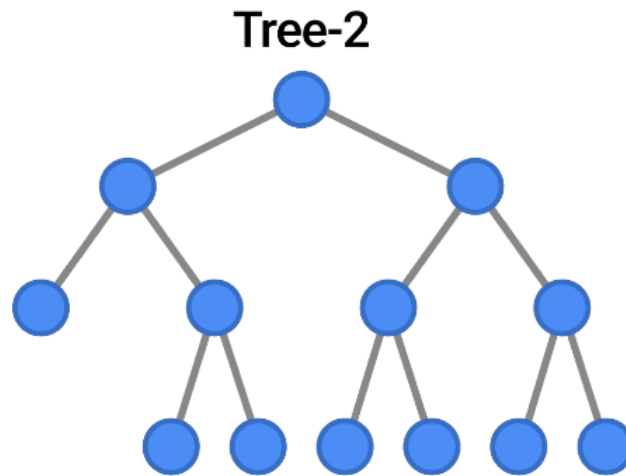
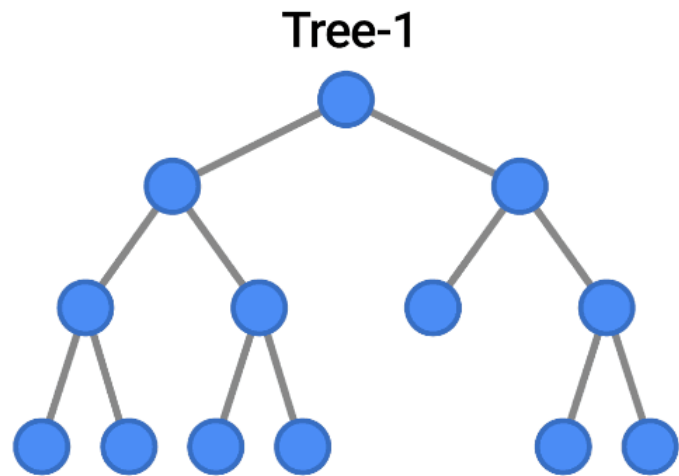
- X : matriz de variables.
- β : coeficientes.
- λ : hiperparámetro de regularización.
- **Ventaja:** Reduce sobreajuste.
- **Desventaja:** Selección de λ no trivial.



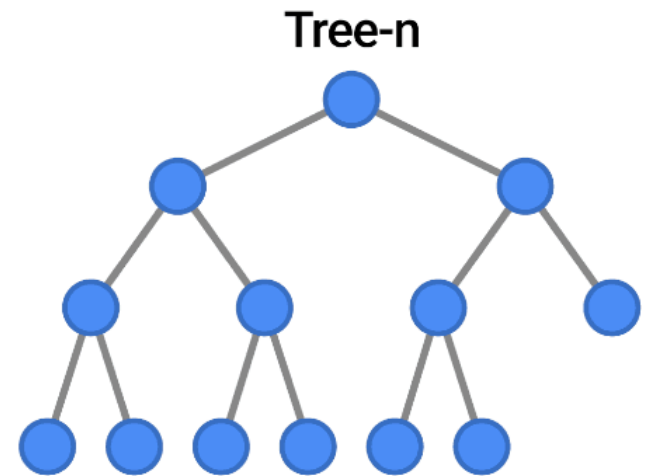
Árboles de Decisión

- **Problemas que resuelve:** Clasificación y Regresión.
- **Fundamento:** División recursiva del espacio de características según impureza (Gini, entropía o MSE).
- **Ventaja:** Muy interpretables.
- **Desventaja:** Inestables con cambios de datos.

EXAMPLES



...



Random Forest

- **Problemas que resuelve:** Clasificación y Regresión.
- **Fundamento:** Conjunto de árboles entrenados con **bagging**.

Idea matemática:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B f_b(x)$$

Random Forest

Nomenclatura:

- B : número de árboles.
- $f_b(x)$: predicción del árbol b .
- **Ventaja:** Generaliza bien, reduce varianza.
- **Desventaja:** Poco interpretable.

Regresión Logística

- Problemas que resuelve: Clasificación binaria.
- Fundamento:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$$

Regresión Logística

Nomenclatura:

- $P(y = 1|x)$: probabilidad de clase positiva.
- β_0 : intercepto.
- β : coeficientes.
- **Ventaja:** Predice probabilidades.
- **Desventaja:** No captura no linealidad compleja.

K-Nearest Neighbors (KNN)

- **Problemas que resuelve:** Clasificación y Regresión.
- **Fundamento:** Asigna la clase/valor según los k vecinos más cercanos.

Distancia Euclídea:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^p (x_{im} - x_{jm})^2}$$

- **Ventaja:** Simple y flexible.
- **Desventaja:** Costoso en datasets grandes.



Support Vector Machines (SVM)

- **Problemas que resuelve:** Clasificación binaria, regresión (SVR).
- **Fundamento:** Encuentra hiperplano con máximo margen.

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{s.a.} \quad y_i(w^T x_i + b) \geq 1$$



Support Vector Machines (SVM)

Nomenclatura:

- w : vector de pesos.
- b : sesgo.
- y_i : etiquetas.
- **Ventaja:** Efectivo en alta dimensión.
- **Desventaja:** Selección de kernel difícil.

Naive Bayes

- **Problemas que resuelve:** Clasificación (texto, spam, sentimiento).
- **Fundamento:** Teorema de Bayes con independencia condicional.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Naive Bayes

Nomenclatura:

- $P(y|x)$: probabilidad posterior.
- $P(x|y)$: verosimilitud.
- $P(y)$: probabilidad a priori.
- $P(x)$: probabilidad marginal.
- **Ventaja:** Rápido, funciona en texto.
- **Desventaja:** Suposición de independencia.

Algoritmos No Supervisados

- **K-Means** (agrupamiento en k clústeres).
- **DBSCAN** (agrupamiento por densidad).
- **PCA** (reducción de dimensionalidad).
- **Clustering Jerárquico** (dendrogramas).

K-Means

- Problemas que resuelve: Agrupamiento.
- Fundamento:

$$\min_{C, \mu} \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

K-Means

Nomenclatura:

- C_i : clúster i .
- μ_i : centroide del clúster i .
- **Ventaja:** Rápido, escalable.
- **Desventaja:** Requiere k fijo.

DBSCAN

- **Problemas que resuelve:** Clustering con ruido.
- **Fundamento:** Puntos denso-conectados con parámetros ϵ , minPts.
- **Ventaja:** Encuentra clusters de forma arbitraria.
- **Desventaja:** Sensible a parámetros.

PCA

- **Problemas que resuelve:** Reducción de dimensionalidad.
- **Fundamento:** Descomposición de la matriz de covarianza.

$$Z = XW$$



Nomenclatura:

- X : datos originales.
- W : autovectores (componentes principales).
- Z : datos transformados.
- **Ventaja:** Reduce ruido y dimensión.
- **Desventaja:** Difícil interpretar componentes.



Clustering Jerárquico


- **Problemas que resuelve:** Agrupamiento con dendrogramas.
- **Fundamento:** Combina clusters según distancia mínima, máxima o promedio.
- **Ventaja:** No requiere k inicial.
- **Desventaja:** Computacionalmente caro en datasets grandes.

Métricas en Regresión

1. MSE (Error Cuadrático Medio)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$


- y_i : valor real.
- \hat{y}_i : predicción.
- n : número de observaciones.

 Evalúa cuánto se alejan las predicciones en promedio, penalizando más los errores grandes.

2. MAE (Error Absoluto Medio)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- y_i : valor real.
- \hat{y}_i : predicción.
- n : número de observaciones.

 Mide la magnitud promedio de los errores sin importar su dirección.

Métricas en Regresión

3. R^2 (Coeficiente de determinación)

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- \bar{y} : media de los valores reales.

 Indica qué proporción de la variabilidad de y es explicada por el modelo (0 a 1).

Métricas en Clasificación

1. Accuracy (Exactitud)

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

 Proporción de predicciones correctas sobre el total.

2. Precision (Precisión)

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

 Indica qué tan confiables son las predicciones positivas (reduce falsos positivos).

Métricas en Clasificación

3. Recall (Sensibilidad o Exhaustividad)

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

 Evalúa la capacidad de detectar verdaderos positivos (reduce falsos negativos).

4. F1-score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

📌 Media armónica entre *Precision* y *Recall*, balancea ambos en un solo valor.