Subject: Probability and Statistics

Class 11: Machine Learning

UdeA

Bioengineering

Francisco José Campuzano Cardona

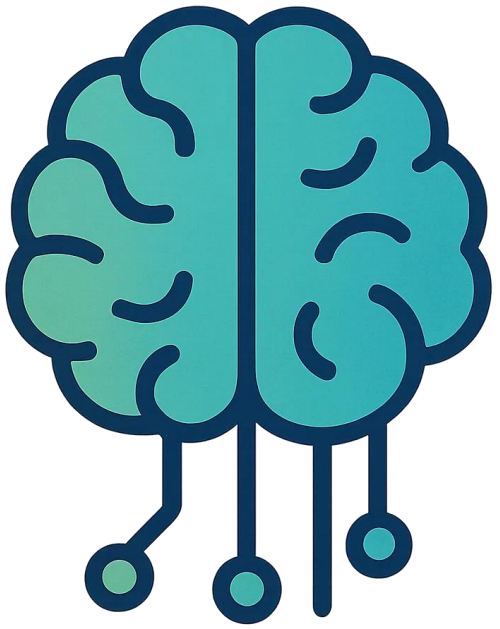
Bioengineerer, MSc in Engineering

Machine Learning – Aprendizaje Automático

UdeA

El aprendizaje automático es clasificado como un subcampo de la Inteligencia Artificial. Y en general, se refiere a la forma **cómo se construyen modelos a partir de los datos.**

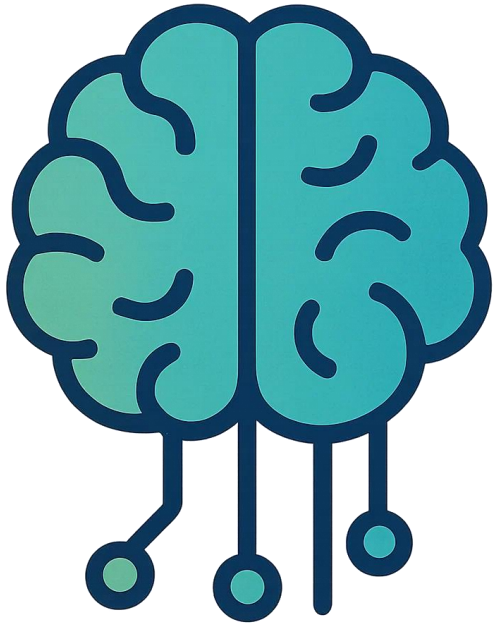
Entonces con el aprendizaje automático básicamente se construyen modelos matemáticos que pueden ayudar a entender los datos. El concepto de aprendizaje aparece porque a estos modelos pueden asignársele parámetros que son **ajustables** en función de los datos, y de este modo se considera que el programa o la máquina **"aprende"** de los datos.



MACHINE LEARNING

UdeA

Los métodos de aprendizaje automático se clasifica en dos categorías



- **Métodos de aprendizaje supervisado**
- **Métodos de aprendizaje no supervisado**

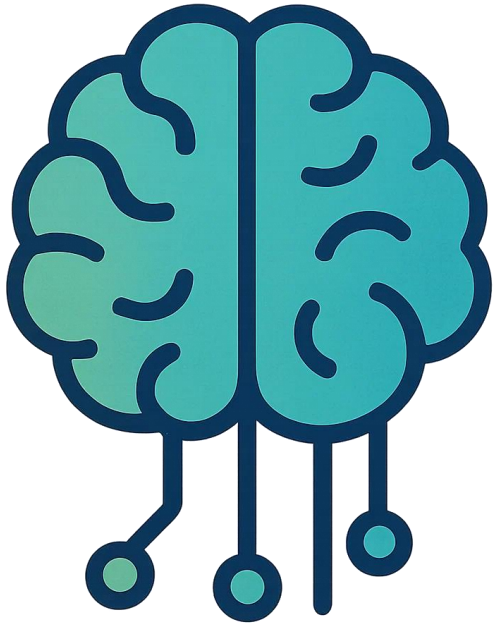
UdeA

Aprendizaje Supervisado

El concepto de supervisado hace referencia a que el método utiliza etiquetas en los datos para entender qué son. Es decir, usa los datos medidos con alguna etiqueta asociada a estos, y **construye o entrena** un modelo

Luego de que se construye el modelo, este puede asignar las etiquetas a los datos desconocidos.

En este tipo de aprendizaje se tienen los **métodos de regresión y métodos de clasificación.**



MACHINE LEARNING

Métodos de regresión: Las etiquetas son variables numéricas continuas.

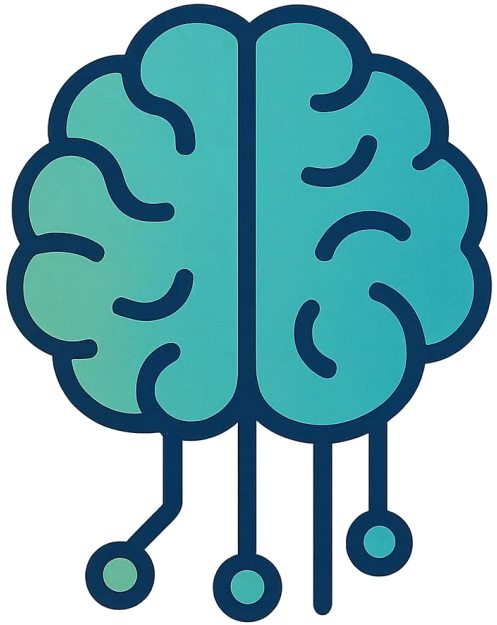
Ejemplo: Usando datos como la edad, el índice de masa corporal (IMC), la presión arterial y la actividad física, se genera un modelo para predecir el valor numérico de **glucosa en sangre (mg/dL) (etiqueta)**

Métodos de clasificación: Las etiquetas son variables numéricas discretas o variables categóricas.

Ejemplo: A partir de atributos como textura, forma y tamaño de células obtenidos de una imagen o muestra, se genera un modelo para clasificar si un tumor es **benigno o maligno (etiqueta)**

UdeA

Aprendizaje No Supervisado



MACHINE LEARNING

En este caso al aprendizaje ser No Supervisado, entonces podemos deducir que no se utilizan etiquetas.

En este tipo de aprendizaje se deja que los datos “hablen por sí solos”. Es decir, en el proceso de construcción o entrenamiento del modelo, no se le “dirá” al modelo, qué es qué.

Dentro de este tipo de aprendizaje se tienen los **Métodos de agrupamiento y los métodos de reducción de dimensión.**

Métodos de Agrupamiento: Los modelos agrupan los datos en grupos homogéneos, según sus características.

Ejemplo: Usando datos de electrocardiogramas (ECG) recogidos de muchos pacientes, se aplica un algoritmo de agrupamiento para identificar grupos de pacientes que tienen patrones similares, sin saber cuántos grupos hay ni cuáles son, por ejemplo, para descubrir subtipos de arritmias o patrones de riesgo que no están previamente clasificados.

Métodos de reducción de dimensión: Los métodos buscan “resumir” los datos. Si tenemos un numero n de variables, buscamos tener un número m, con m<n, sin perder información.

Ejemplo: En un estudio de expresión genética, cada muestra puede tener miles de genes medidos. Usamos reducción de dimensión para condensar toda esa información en unas pocas variables que expliquen la mayor parte de la variabilidad

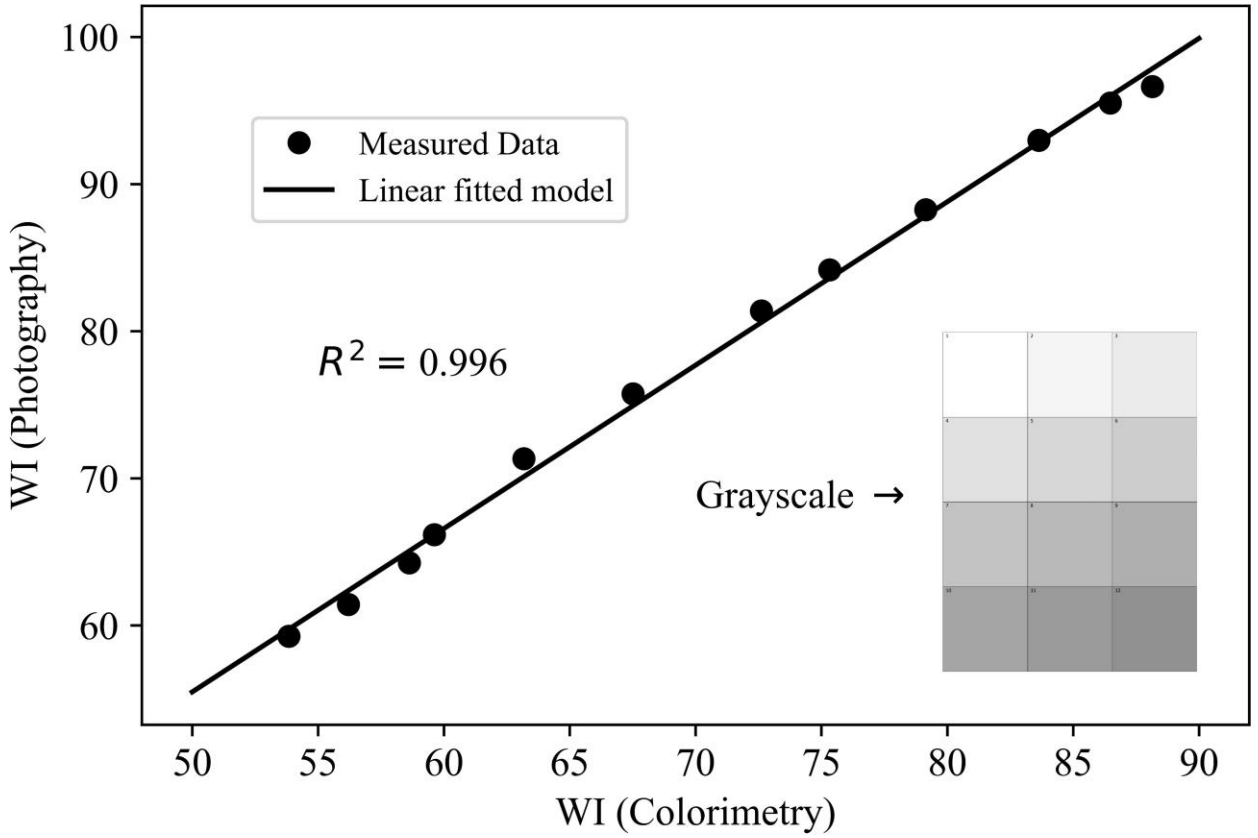
UdeA

Regresión Lineal Simple

En este caso se modela una variable ***Y*** en términos de otra variable, suponiendo que su relación es proporcional, y por tanto el modelo será el de una línea recta.

$$Y = b_0 + b_1 X$$

Siendo ***X*** la variable descriptora.



UdeA

Key Terms for Simple Linear Regression

Response

The variable we are trying to predict.

Synonyms

dependent variable, Y-variable, target, outcome

Independent variable

The variable used to predict the response.

Synonyms

independent variable, X-variable, feature, attribute

UdeA

Record

The vector of predictor and outcome values for a specific individual or case.

Synonyms

row, case, instance, example

Intercept

The intercept of the regression line—that is, the predicted value when $X = 0$.

Synonyms

b_0, β_0

Regression coefficient

The slope of the regression line.

Synonyms

slope, b_1, β_1, parameter estimates, weights

UdeA

Fitted values

The estimates $\hat{Y}_i$ obtained from the regression line.

Synonyms

predicted values

Residuals

The difference between the observed values and the fitted values.

Synonyms

errors

UdeA

Regresión Lineal Simple

Este método es supervisado, ya que la construcción el modelo necesita un set de datos pareados de $\boldsymbol{X}$ y $\boldsymbol{Y}$ para la construcción del modelo. En este caso los $\boldsymbol{Y}$ medidos, son las etiquetas.

El método encuentra un intercepto y una pendiente que son estimadores de los parámetros reales.

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X$$

Pero ¿Cómo se encuentran estos estimadores?

UdeA

Mínimos Cuadrados o Mínimos cuadrados ordinarios.
Least Squares or Ordinary Least Squares (OLS)

El modelo se genera buscando minimizar el error entre los valores predichos y los valores medidos. Esto se hace minimizando la suma de los errores, pero como estos podrían ser algunos positivos y otros negativos, y en consecuencia anularse, se toman los cuadrados, y por esto el método se llama el "Mínimos Cuadrados"

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X}$$

UdeA

Predicción Vs. Explicación

Según el contexto, la regresión puede usarse para predecir una variable Y o para explicarla.

Por ejemplo, en salud pública, se podría estar interesado en determinar si una campaña informativa causa algún impacto sobre el uso de prácticas sexuales seguras. En este caso, se está explicando una variable por otra, pero no se busca predecir.

En otro caso, por ejemplo, se determina la relación entre el IMC y la presión arterial, se podría usar el modelo para predecir si alguien es hipertenso conociendo su IMC.

UdeA

Causalidad

Un concepto importante en la regresión, es la causalidad. La regresión por si sola no es capaz de determinar la causalidad. Es decir, no puede afirmar si Y es causado por X o si X es causado por Y.

El entendimiento de la causalidad depende del entendimiento del contexto de las variables, y de los fenómenos que describen, por tanto responsabilidad del analista, y no del método en sí.

En el ejemplo de la hipertensión y el IMC, en términos matemáticos, se podría decir que la hipertensión causa obesidad, pero sabemos que no es la dirección de causalidad correcta.

UdeA

Regresión Lineal Múltiple.

La regresión lineal se puede generalizar a varias variables predictoras. Es decir, una variable $\boldsymbol{Y}$ puede ser predicha por un modelo que combina el peso de múltiples variables $\boldsymbol{X_i}$

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_i X_i$$

El modelo continua siendo lineal, dado que Y se explica por la combinación lineal de las variables $\boldsymbol{X_i}$

UdeA

¿Cómo evaluamos la calidad de un modelo de regresión?

- **Error de raíz cuadrada media (RMSE):** Es una medida de la precisión general del modelo, y sirve de comparación con otros modelos, el modelo con menor RMSE es más preciso.

- **Error residual estándar (RSE):** Al igual que RMSE es una medida de precisión general y sirve de comparación con otro modelo, a menor RSE, mejor precisión tiene el modelo. (p # de predictores)

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

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n - p - 1)}}$$

UdeA

¿Cómo evaluamos la calidad de un modelo de regresión?

- **Error Absoluto medio (MAE):** Es una medida de la precisión general del modelo, y sirve de comparación con otros modelos, el modelo con menor MAE es más preciso. Tiene de ventaja que es robusto ante datos atípicos.

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

En regresión, un dato atípico, es un dato con error muy grande, es decir, un valor predicho que se aleja mucho del valor medido.

UdeA

¿Cómo evaluamos la calidad de un modelo de regresión?

- **Coeficiente de determinación - R^2**. Esta métrica se encuentra entre 0 y 1, mide la proporción de variación tenida en cuenta por el modelo, en otras palabras, indica qué tan bien se ajustan los datos a una relación lineal, donde 1 es una relación perfecta, y 0 una relación nula.

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

No existe una regla para decir qué valor de R^2 es apropiado, depende de la variabilidad del sistema que se esté modelado. Pero en general a mayor R^2 mejor, y valores inferiores a 0.6 no deberían ser aceptados.

UdeA

- **Validación cruzada:** Se refiere a probar el modelo con datos que no fueron tenidos en cuenta para el entrenamiento del modelo. Es decir, se dejan cierta cantidad de datos aparte, y se usan para probar el desempeño del modelo, usando por ejemplo RMSE, RSE, o R^2

Pero la submuestra dejada aparte, podría tener cierta variabilidad propia, ¿Cómo tenerlo en cuenta?

- ***k-folds cross-validation.***

UdeA

K-folds cross-validation. (k entre 5 y 20)

1. Se toma 1/k datos de la muestra, y se entrena el modelo sin estos datos. Luego se evaluar el desemepeño del modelo usando los datos que fueron dejados a parte
2. Se realiza otro submuestreo 1/k, asegurándose de no dejar aparte los datos que se dejaron ya por fuer del modelo. Se entrena el modelo con los datos restantes y se evalúa el desempeño con los datos dejados aparte.
3. Se repite hasta usar todos los datos del set para probar
4. Finalmente se promedian los valores de desempeño.

UdeA

Otras validaciones importantes.

Además de las validaciones de precisión, un modelo de regresión apropiado debería entregar residuos no autocorrelacionados, de lo contrario habría una mala especificación del modelo.

Para corroborar esto se puede usar el estadístico Durbin-Watson, el cual indica autocorrelación si está cercano a 0 o 4, y no autocorrelación si está en cercano a 2.
La normalidad de los residuos también es un requisito.

Sin embargo, en los modelos de regresión, el error aceptado depende de lo que se esté modelando.

Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering