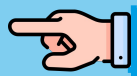
Aprendizaje Supervisado - Regresión

Clase sincrónica

{desafío} latam_

Implementar modelos de aprendizaje automático por medio de técnicas estadísticas, adecuando los diferentes algoritmos debidamente a la situación y requerimientos necesarios

- Unidad 1: Introducción al Machine Learning
- Unidad 2: Aprendizaje Supervisado y No Supervisado
  (Parte I: No supervisado)
  (Parte II: Clasificación)
  (Parte III: Clasificación)
  (Parte IV: Regresión)
  (Parte V: Series de tiempo)
- Unidad 3: Aplicando lo aprendido
  (Parte I: Preprocesamiento de datos)
  (Parte II: Modelamiento)



{desafío} latam_

¿Qué aprenderás en esta sesión?

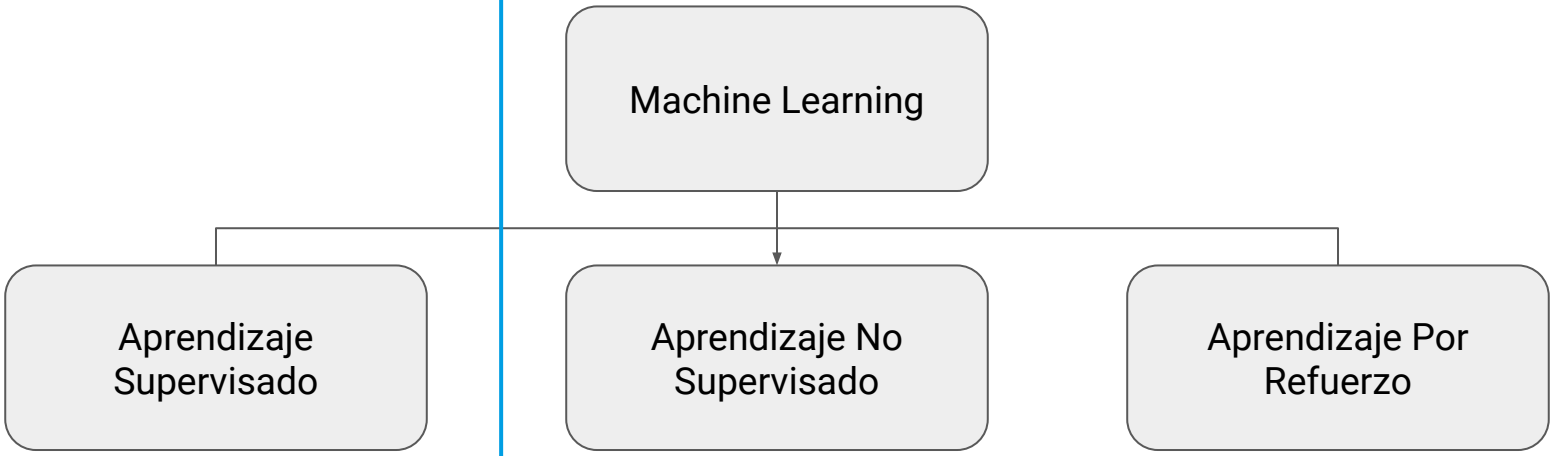
En esta sesión aprenderás sobre los algoritmos de aprendizaje supervisado, específicamente sobre algoritmos de regresión. Al finalizar sabrán para qué sirven y cómo implementar los algoritmos.

{desafío}
latam_

¿Donde se ubica la regresión en las tareas de machine learning?

{desafío} latam_

Aprendizaje Supervisado



El aprendizaje supervisado se refiere a la capacitación de modelos utilizando datos de entrenamiento etiquetados.

- Clasificación
- **Regresión**

¿A qué se refieren los conceptos de overfitting y underfitting?

{desafío} latam_

¿Qué función tiene el kernel en el support vector machine?

{desafío} latam_

/* Regresión */

Regresión

¿En qué consiste?

En este caso, tenemos una variable objetivo numérica, se intenta entender el comportamiento entre las variables predictoras y un valor continuo.

Variables Predictoras: Se utilizan para predecir la variable objetivo

age	sex	bmi	bp	glucosa
0.038076	0.050680	0.061696	0.021872	-0.017646
-0.001882	-0.044642	-0.051474	-0.026328	-0.092204
0.085299	0.050680	0.044451	-0.005670	-0.025930
-0.089063	-0.044642	-0.011595	-0.036656	-0.009362
0.005383	-0.044642	-0.036385	0.021872	-0.046641
...	...	...	...	...
0.041708	0.050680	0.019662	0.059744	0.007207
-0.005515	0.050680	-0.015906	-0.067642	0.044485
0.041708	0.050680	-0.015906	0.017293	0.015491
-0.045472	-0.044642	0.039062	0.001215	-0.025930
-0.045472	-0.044642	-0.073030	-0.081413	0.003064

Variable Objetivo: numérico (Valores continuos con cierta distribución)

Regresión

Aplicaciones

- **Economía**: Estimar el PIB de un país o región utilizando variables económicas como el consumo, la inversión y el gasto público.
- **Medicina:** Predecir la tasa de crecimiento de un tumor basándose en características médicas y datos de pacientes.
- **Agricultura**: Estimar el rendimiento de los cultivos basándose en datos climáticos, de suelo y de cultivo.
- **Manufactura**: Predecir el tiempo de vida útil de un componente o maquinaria basándose en datos de mantenimiento y uso.

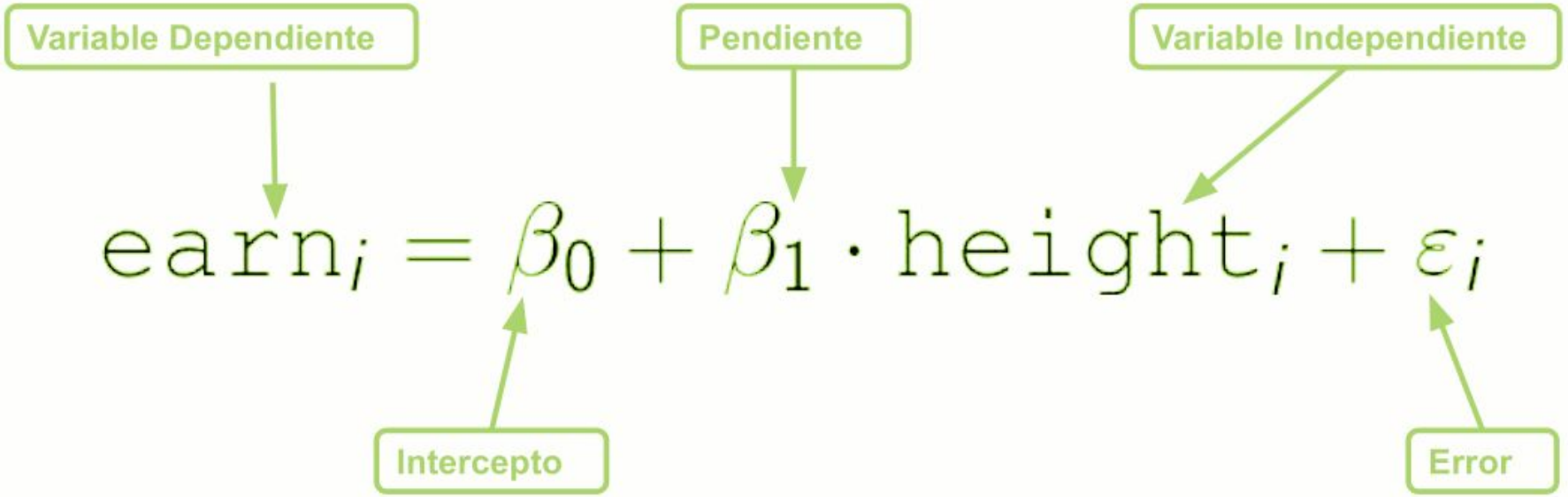
Regresión

Aplicaciones

- **Marketing:** Predecir las ventas futuras de un producto basándose en datos de marketing, promociones y precios.
- **Seguros:** Predecir el costo de las reclamaciones basándose en datos de seguros y características del asegurado.
- **Medio Ambiente:** Estimar la concentración de contaminantes atmosféricos basándose en datos de calidad del aire y factores ambientales.

/* Regresión Lineal */

Regresión Lineal



Estimación de parámetros en regresión lineal

Mínimos cuadrados

Para estimar los betas de la regresión lineal (“entrenar el modelo”) se utiliza el método de Mínimos cuadrados ordinarios (MCO o OLS), con el cual se busca ajustar los betas al mínimo error.

$$\beta = argmin\mathrm{E}[(y_i - X^T\beta)^2]$$

$$\beta = \sum(y_i - (\beta_0 + \beta_1 x)^2)$$

Regresión Lineal

Ventajas y desventajas

Ventajas

- **Interpretabilidad**: es fácil de interpretar, lo que facilita la comprensión y explicación del modelo.
- **Eficiencia y Velocidad:** es computacionalmente eficiente.
- **Implementación Sencilla**: fácil de implementar y entender.

Desventajas

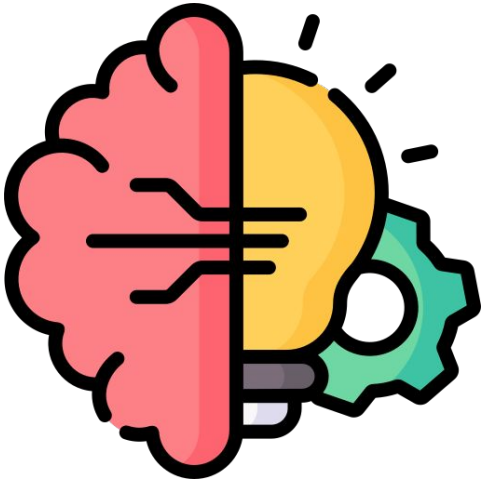
- **Sensibilidad a Outliers**: es sensible a los valores atípicos.
- **Limitaciones en Datos No Lineales:** no se ajusta a todos los modelos.
- **Multicolinealidad**: si las características están altamente correlacionadas entre sí, la regresión lineal puede producir estimaciones inestables.

Enfoque Machine Learning

¿Qué lo caracteriza?

Recordemos que estamos observando la regresión lineal desde el enfoque de Machine Learning.

Esto significa que nuestro objetivo es poder estimar de la mejor forma en nuevos datos, es decir, poder generalizar el comportamiento del modelo.

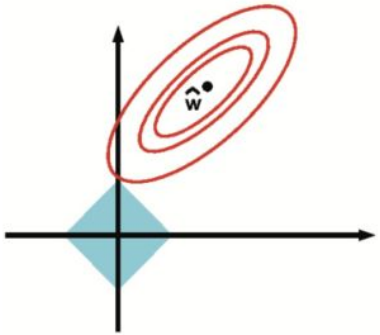


/*Regularización*/

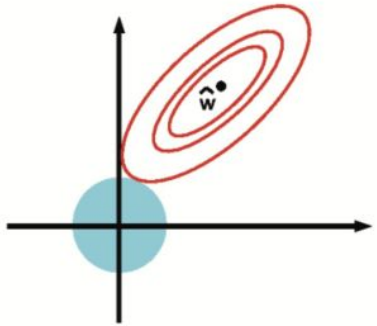
Regularización

Definición y normas

La regularización es una técnica utilizada para controlar y evitar el sobreajuste (overfitting) del modelo. Se implementa utilizando normas para penalizar los parámetros.



Norma L1 (Lasso): Se mide la distancia entre 2 vectores según la norma absoluta.



Norma L2 (Ridge): Sintetiza la distancia entre dos vectores mediante la norma euclídea.

Ridge

Características y cálculo

$$\beta_{\text{Ridge}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Ridge modifica la superficie de penalización de los coeficientes mediante el **hiperparámetro lambda**.
- **Lambda** gobierna la **superficie de penalización** que está determinada por la cantidad de parámetros inferidos en el modelo.
- Dado que tiene una forma cuadrática, **suaviza pero no elimina** atributos irrelevantes.

Lasso

Características y cálculo

$$\beta_{\text{Lasso}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Principal diferencia con Ridge: permite seleccionar y **eliminar atributos** irrelevantes del modelo.
- De igual manera que en Ridge, el **hiperparámetro lambda** define el área de la superficie de penalización.
- La diferencia radica en la **norma de penalización.**

Elastic Net

Características y cálculo

$$\beta_{\text{ElasticNet}} = \underset{\beta}{\text{argmin}} \sum_{i}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

- Elastic Net combina ambas normas de penalización.
- L1 nos asegura una selección de atributos.
- L2 nos asegura una penalización parsimoniosa de los coeficientes de los atributos.
- Existe un parámetro que gobierna la dominancia entre ambas formas de penalización.

¡Manos a la obra!
Probando las diferentes penalizaciones

¡Manos a la obra!

Penalizaciones con Python

Veremos a continuación la implementación de las penalizaciones con Python. Para esto, observa los pasos que mostrará tu profesor en la presentación de Jupyter Notebook (puedes abrir el archivo adjunto de tu guía para observar y replicar los pasos). En esta presentación veremos:

Regularización en regresión lineal:

a. Ridge regression
b. Lasso regression
c. Elastic Net

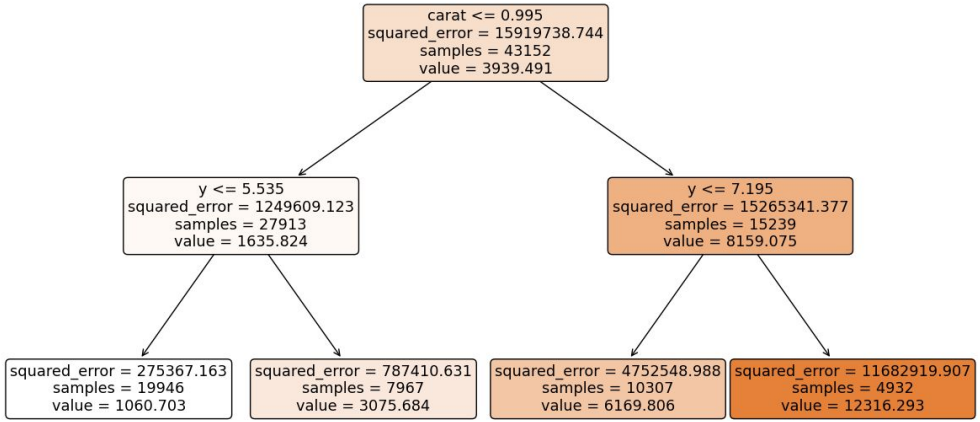
{desafío} latam_

/* Árboles de regresión */

Árboles de regresión

¿En qué consiste?

Son una extensión de los árboles de clasificación que predicen un valor numérico. En vez de utilizar criterios de pureza, utilizan un métricas de regresión.

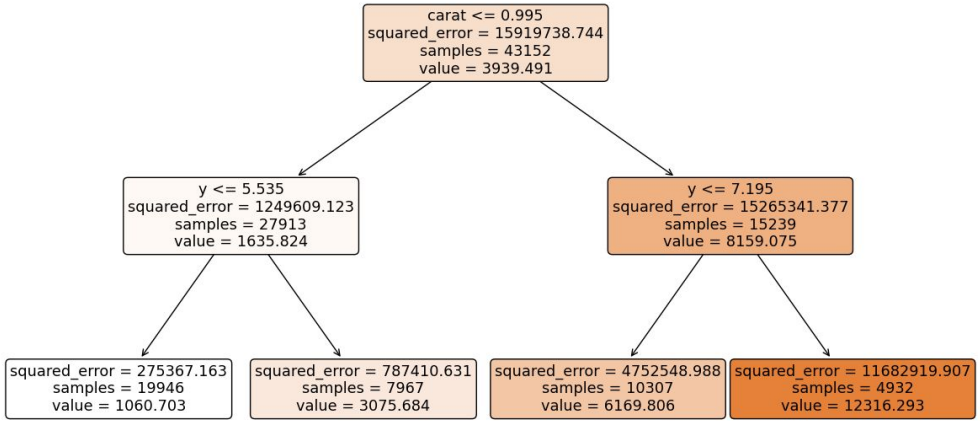


$$MSE = \frac{1}{n}\sum(y - \hat{y})^2$$

Árboles de regresión

¿Cómo funcionan?

Los árboles de regresión funcionan de manera similar a los árboles de clasificación, pero en lugar de predecir una etiqueta de clase mayoritaria en cada nodo hoja, se predice un valor numérico promedio basado en las muestras que llegan a esa hoja.



Árboles de regresión

Construcción

Selección de la división	Cálculo de la predicción	Criterio de división	Crecimiento del árbol
El árbol comienza con un nodo raíz que contiene todos los datos de entrenamiento. En cada paso, se selecciona una característica y un umbral que dividirá los datos en dos grupos.	Se calcula la predicción numérica para cada región basada en los valores promedio de las muestras en esa región.	La elección de la característica y el umbral se realiza de manera que la reducción en el error de predicción sea máxima después de la división. El error se mide en términos de alguna métrica, como la suma de los cuadrados de los residuos (SSE) o la desviación absoluta media (MAD).	El proceso de selección y división se repite para cada región creada en pasos anteriores, hasta que se cumple algún criterio de detención, como la profundidad máxima del árbol o el número mínimo de muestras en una región.

Árboles de regresión

Diferencias con la clasificación

	Regresión	Clasificación
Tipo de salida	Discreta	Continua
Métricas de evaluación	MSE - RMSE	Precision - Recall - F1-score
Función de división	Reducción del error en términos de precisión numérica	Búsqueda de pureza en la clasificación

Árboles de regresión

Ventajas

Interpretabilidad

No linealidad

Robustez ante outliers

Flexibilidad

Tratamiento automático de variables

Árboles de regresión

Desventajas

Sobreajuste

Estabilidad

Limitaciones con la extrapolación

No considera relaciones globales

Árboles de regresión

Hiperparámetros

Hiperparámetro	Descripción
Máximo de Profundidad	hasta qué punto puede crecer
Cantidad de atributos	cuántos atributos se deben considerar
Mínimo de muestras en un nodo particionable	con cuántas observaciones podemos seguir subdividiendo.
Mínimo de muestras en un nodo terminal	con cuántas observaciones podemos dejar de subdividir.
Criterio de split	criterio para generar la mejor división.

¡Manos a la obra!
Árboles con Python

{desafío}
latam_

¡Manos a la obra!

Árboles con Python

En la tutoría verás cómo implementar árboles de regresión con Python. Específicamente, podrás ver::

1. Construcción de un árbol de regresión
2. Ajuste de hiperparámetros

{desafío} latam_

/* Métricas de regresión */

Métricas de regresión

Evaluando el modelo

Como el objetivo de ML es poder generalizar el modelo a nuevos datos es importante poder medir el **error del modelo,** Para esto tenemos múltiples métricas a disposición como:

1. **MAE:** Mean Absolute Error
2. **MSE:** Mean Square Error
3. **MAPE:** Mean Absolute Percentage Error
4. **Otros varios**

Métricas de regresión

Fórmulas

Métrica	Fórmula
Error Cuadrático Medio (MSE)	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Raíz del Error Cuadrático Medio (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Desviación Absoluta Media (MAE)	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Error Absoluto Porcentual Medio (MAPE)	$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100\%$
Coeficiente de determinación (R2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Ideas fuerza

Los problemas de ***Regresión*** *corresponden a la predicción de variables objetivo continuas..*

Muchos de los modelos que se usan para clasificación tienen su adaptación a problemas de regresión como árboles de regresión.

Hay múltiples métricas para problemas de regresión, donde las más utilizadas son ***R2, RMSE, MAPE, etc****.*

¿Qué conceptos no te quedaron claros o quieres reforzar?

Recursos asincrónicos

¡No olvides revisarlos!

Para esta semana deberás revisar:

- Guía de estudio
- Desafío “Prediciendo el precio de las casas”

Próxima sesión...

Veremos modelos de series temporales, los usos e implementación en python.

{desafío}
latam_

Academia de
talentos digitales