**Regresión Lineal**

La regresión lineal es una técnica que nos ayuda a hacer comparaciones entre los puntos en el plano cartesiano. es decir, que para un valor de X existe un valor de Y, esto se empieza a graficar sobre el plano como un gráfico de dispersión (*scatter plot*). Hace parte de los modelos de machine learning de Análisis supervisado.

Para esos datos, hay que ajustar una linea recta que mejor los entienda, ósea el modelo trata de predecir el mejor ajuste para la linea recta en esos datos.

La función matemática correspondiente a este problema es Y=w0 + w1x

* X es un valor que existe sobre el plano
* W0 es el punto que se esta cortando sobre el plano, es decir, cuando X = 0 cual es el salto que existe el punto de X = 0 y el plano en el eje Y.
* W1 es cuando se agrega un valor a la X cuanto salto sobre Y

Básicamente esto se conoce como w0 el Intercepto y w1 la pendiente de la linea recta.

durante el ajuste del modelo de regresión lineal se conocen como los pesos del modelo, estos son los valores que va a estar intentando aprender el modelo para entender cual va a ser la mejor linea recta que se va a ajustar a los datos.

**¿Cuándo utilizar el modelamiento lineal?**

1. ¿Tengo que predecir sobre una variable numérica?

Si mide algo, entonces check.

1. ¿Las variables independientes con las que cuento son primordialmente numéricas?

Si lo son, es otro check.

1. No cuento con una gran cantidad de variables y/o variables categóricas con muchos niveles. Otro check.

**Recomendaciones**

* Si tengo que predecir variables numéricas empiezo con la regresión lineal, si no funciona, salto a otros modelos.
* Reduce las variables (dimensionalidad) lo más que se pueda
* ¡MULTICOLINEALIDAD! NO tomar variables que estén muy relacionadas entre sí.

Ejemplo: Si dentro del dataset hay peso en kg y peso en lbs son esencialmente lo mismo, no deberían usarse. Las que están relacionadas, sería mejor usar sólo una.

* No puedo predecir fuera del dominio de la variable independiente.

Ejemplo: si van los datos en x de 0 a 20, sólo usarlo entre ese rango de datos. Podría ser muy distinto más arriba o más abajo.

**Función de pérdida y optimización: mínimos cuadrados:**

Un algoritmo de Machine Learning siempre pasa por un loop:

1. Se ajusta el modelo.

Ejemplo: Crear la línea recta en la regresión lineal.

1. Se comparan resultados con los datos reales.
2. Se ajustan pesos en el modelo.
3. Regreso a paso inicial si no se converge.

**Función de pérdida:**

Es el estándar que se utiliza para poder saber cuál es el porcentaje de acierto del modelamiento con respecto a los datos reales. Lo que se busca justamente es maximizar o minimizar según sea el caso. En el caso de una regresión lineal es el método de mínimos cuadrados el que se usa más.

MSE Mean Square Error

R^2 Coeficiente de determinación. Observa qué tan bien el modelo se ajusta según la varianza del los datos. Mientras más se acerca a 1 es mejor.

Otras menos usadas:

R-Ajustada es una variación al R^2 para cuando hay muchas variables dentro del modelo

Error máximo

MAE Error absoluto promedio

Mediana de los errores absolutos

RMSE Raíz del promedio de los errores cuadrados

Varianza explítica

GRÁFICA DE RESIDUALES PARA GIRAR LA REGRESIÓN A PENDIENTE INDETERMINADA (LÍNEA PARALELA A X) PARA VER EL ERROR REAL MÁS GRÁFICAMENTE.

**Algoritmo de optimización:**

Descenso del gradiente. Girar la recta en su pendiente y corte para entender mejor el gradiente de cambio.

**REGRESIÓN LINEAL MULTIVARIADA**

Y = w0 + w1\*x1 + w2\*x2 + … wn\*xn