

# Machine Learning

- Procesamiento de datos
- Regresión lineal simple
- Regresión Lineal Multiple
- Regresión Polinómica
- Regresión con máquina de soporte vectorial SVM
- Árboles de decisión para Regresión
- Regresión con bosques aleatorios

# Procesamiento de datos

## Pre-procesado de datos

### Introducción

Tenemos una serie de datos, ya observados, la idea es suministrar estos datos al machine-learning y la máquina va a **intentar definir una variable dependiente**.

### Notas Python y R

- Cambiar formato decimales en spider por 0f, si no obtendremos los números en anotación científica.
- Python inicia a contar desde 0
- R inicia a contar desde 1

## Variables Machine Learning

- Variables independientes → Son las variables que le daremos al algoritmo para intentar predecir.
- Variables dependientes → Es la variable que queremos predecir.

## Datos Desconocidos

Cuando nos encontramos con ausencia de valores podemos optar por introducir la media o la moda de dicha columna, siempre será mejor esto que poner ese valor a 0.

## Datos Categóricos

Se tratan de esos valores que su columna en vez de tener un número contiene un valor para catalogar o clasificar un usuario.

Variable dummy → traducir una variable a variable categorica sin orden. La variable dummy clasificada en activa

# Set de entrenamiento y Set de Test

Over fitting → problema que hay que intentar evitar. El algoritmo no tiene comparaciones suficientes y aprende lo

- 70% o 80% para entrenamiento
- 20% o 30% para testing

## Escalado de datos

Diferencias de rango de valores, ejemplo edad(27) y salarios(51000). El efecto de la edad pasaría inadvertido en nuestro algoritmo de machine learning.

Si tenemos una variable cuyo rango de valor es muy superior a las otras, podría ser un problema porque las varia

- Normalización de valores menor valor -1 mayor valor 1: Standardización o Normalización
- Standardización → Permite aglutinar valores en torno a la media
- Normalización → El más pequeño es 0 el mayor es 1

[Normalizacion.png](#) unknown

# Regresión lineal simple

Vamos a intentar predecir y crear un modelo lineal, regresión lineal simple. Buscará todas las rectas posibles y nos dirá cuál es la recta que más se acerca la distancia de la propia recta a los puntos de referencia.

Es la línea de tendencia que más se ajusta a los datos ofrecidos.

## Variables

### Categóricas

- Nominales -> Rojo,verde,azul,... (Factores)
- Ordinales -> Pequeño,Mediano,Grande, A,B,C (Tiene un orden)

### Numéricas

- Discretas -> 800 empleados (objetos que podemos contar sin usar decimales)
- Continuas -> El peso, la altura, entran todo tipo de números.

## La palabra regresión

LLamamos análisis de regresión al prece estadístico de estimar las relaciones que existen entre variables.

Se centra en estudiar las relaciones entre una variable dependiente de una o varias variables independientes.

[Regresión.png](#) type unknown

## Regresión Lineal Simple

[Regresión lineal.png](#) type unknown

Lo que hará nuestro algoritmo de regresión lineal es sumar todas las diferencias, las rectas entre  $y_i$  y  $\hat{y}_i$ , las elevará al cuadrado porque algunas serán positivas y otras negativas. De todas las rectas se que con aquella que minimiza los cuadrados de las diferencias entre el dato real y la predicción.

# Regresión Lineal Multiple

## Restricciones de la Regresión Lineal

1. Linealidad
2. Homocedasticidad
3. Normalidad Multivariable
4. Independencia de los errores
5. Ausencia de multicolinealidad. El modelo es incapaz de distinguir los efectos de una variable dummy.

## Variable Dummy

Cuando necesitemos construir un modelo con variables ficticias (variables dummy), hay que omitir uno de los factores, uno de los niveles de la variable ficticia. Es decir, si tenemos 100 países como variables ficticias, solo colocaríamos como dummy 99 en el modelo de regresión.

## P-Valor

### Que no es

- El p-valor no es la probabilidad de que la afirmación sea cierta.
- El p-valor no es la probabilidad de que la hipótesis nula sea cierta.

Nos indica que tan probable es obtener un resultado con una hipótesis nula verdadera.

[\[\[1\]\]](#)

## Paso a paso en la regresión lineal multiple

# Añadir todas las variables independientes

No, por dos razones:

1. NO por añadir más variables vamos a tener más información. In Basura = Out Basura.
2. Si el número de variables va creciendo hace difícil la explicación lógica del proceso.

## 5 métodos obtención de variables relevantes

Tenemos 5 métodos disponibles para obtener las variables importantes y que tendrán relevancia en el algoritmo de regresión lineal múltiple.

### Exhaustivo (All-in)

Metemos todas las variables en el modelo, razones por las que hacer esto:

- Conocimiento previo de todas las variables. Todas son variables predictoras.
- Por necesidad, nos obligan a utilizar todas las variables.
- Preparación previa para realizar la eliminación hacia atrás.

### Eliminación hacia atrás

1. Selección del nivel de significación en el modelo, normalmente  $SL=0.05$
2. Se calcula el modelo con todas las variables
3. Se obtiene la variable predictoras con el p-valor más grande. Si  $P > SL$ , entonces pasamos al paso 4, sino vamos a fin.
4. Se elimina la variable predictora.
5. Reajuste del modelo sin dicha variable.

Con el nuevo modelo creado, las variables de ese tendrán una serie de p-valores y por tanto repetimos el paso 3,

### Selección hacia adelante

1. Seleccionamos un nivel de significación, pero en este caso será para entrar en el modelo.
2. Ajustamos todos los modelos de regresión lineal simple. Elegimos el que tiene el menor p-valor.

3. Conservamos estas variables y ajustamos todos los modelos con la variable extra añadida a la que ya tenga el modelo en ese momento.
4. Consideramos la variable predictora con el menor p-valor. Si  $P < SL$  volvemos al paso 3.

Seguiremos sucesivamente añadiendo variables mientras el p-valor sean inferior al nivel de significación, en el caso contrario se detiene el proceso.

## Eliminación bidireccional o regresión dual

1. Seleccionamos dos niveles de significación para entrar y salir del modelo.
2. Selección hacia delante  $p\text{-valor} < SL_{\text{Enter}}$
3. Selección hacia atrás  $p\text{-valor} < SL_{\text{Stay}}$
4. No hay nuevas variables para entrar ni tampoco variables antiguas para salir

## Comparaciones de puntos

1. Seleccionar un criterio de la bondad de ajuste. Cuando un modelo será mejor que otro.
2. Construir todos los Modelos:  $2^N - 1$
3. Selección del modelo con mejor criterio

OJO! -> 10 columnas de datos = 1023 modelos



# Regresión Polinómica

# Regresión con máquina de soporte vectorial SVM

# Regresión con máquina de soporte vectorial SVR

Sirven tanto para regresiones lineales como no lineales. La idea es ajustar una calle, he intentar mantener cuántas más obervaciones posibles del conjunto de datos dentro de la calle, limitando unos márgenes máximos.

## Hyper parámetro épsilon

La anchura del pasillo se controla mediante un hiper parámetro, épsilon. Cuánto mayor es ese valor, mayor es la anchura de la calle.

## Objetivo

En la regresión lineal se intenta minimizar el error entre la predicción y los datos. En SVR el objetivo es que los err

# Arboles de decisión para Regresión

## Árboles de regresión

CART -> Classification and Regresión Tree

Una vez ejecutamos nuestro algoritmo árbol de decisión, el conjunto de datos de las variables independientes quedará dividido en segmentos.

[Arbol de decision.png](#)  


Básicamente se fija en la entropia de los puntos para poder agruparlos, cada una de estas divisiones aporta una información realmente buena.

[Arbol decision2.png](#)  


Podemos ver en verda la media de los segmentos y en la imagen de abajo el árbol de decisiones

[Arbol decision3.png](#)  


# Regresión con bosques aleatorios

## Bosques aleatorios

Versión mejorada del árbol de regresión ya que es capaz de utilizar miles de árboles de regresión para obtener una mejor predicción.

## Pasos a seguir

1. Elegir un número aleatorio  $K$  de puntos de datos del conjunto de Entrenamiento.
2. Árbol de desición asociado a esos  $K$  puntos.
3. Elegir el número de NTree de árboles que queremos construir y repetimos Paso 1 y Paso 2
4. Cada uno de los árboles hace una predicción del valor  $Y$ , luego hacemos un promedio de esos NTree predicciones.