## Regresión lineal y múltiple

El algoritmo de regresión lineal nos ayuda a conseguir tendencia en los datos, este es un algoritmo de tipo supervisado ya que debemos de usar datos previamente etiquetados.

En la regresión lineal generamos, a partir de los datos, una recta y es a partir de esta que podremos encontrar la tendencia o predicción.

Generalmente es importante tener en cuenta varias dimensiones o variables al considerar los datos que estamos suministrando al modelo, recordando siempre cuidar este set de sobreajuste o subajuste.

Cuando nuestro modelo considera más de dos variables el algoritmo de regresión que usamos se conoce como Regresión Lineal Múltiple y este trabaja sobre un sistema de referencia conocido como hiperplano.

Los algoritmos de regresión, tanto lineal como múltiple trabajan únicamente con datos de tipo cuantitativos.

**Sobreajunte (overfiting)**: Es cuando intentamos obligar a nuestro algoritmo a que se ajuste demasiado a todos los datos posibles. Es muy importante proveer con información abundante a nuestro modelo pero también esta debe ser lo suficientemente variada para que nuestro algoritmo pueda generalizar lo aprendido.

Todos gatos, pero solo datos de gatos blancos, y cuando viene uno negro no lo identifica.

**Subajuste (underfiting)**: Es cuando le suministramo a nuestro modelo un conjunto de datos es muy pequeño, en este caso nuestro modelo no sera capas de aprender lo suficiente ya que tiene muy poca infomación. La recomendación cuando se tienen muy pocos datos es usar el 70% de los datos para que el algoritmo aprenda y usar el resto para entrenamiento.

El set de gatos muy pequeno.

X: matriz x:vector subconjunto de matriz

X:Poblacion x:variable aleatoria de la poblacion

Muchas variables X1, X2, …Xn la información mas trascendente es en pocas variables. Se pueden evaluar con

PCA (principal component analysis)

El árbol de decisión implementado en WEKA

‘Feature selection’ \*\*(o selección de variables) junto con la implementación del PCA, sklearn lo tiene: [Feature Selection](https://scikit-learn.org/stable/modules/feature_selection.html)

 eigenvectores del PCA (los componentes principales) **corresponden a una combinación lineal de TODAS variables** en la que cada una aporta con un cierto ‘peso’ (loadings), las variables más importantes (que más varianza aportan al componente principal en cuestión) son aquellas con loadings más grandes. Así que, en general, uno busca conservar las variables con loading más altos en los Componentes Principales (CP) que aportan más a la varianza total de los datos (explained variance).