UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

# Introducción al Aprendizaje Automático

Modelos de regresión: lineal

Facultad de Ingeniería

Universidad de la República

# Aprendizaje Supervisado

- Tenemos un conjunto de datos {<x,y>}, formado por instancias x de un dominio cualquiera D, cada una de ellas con un valor asociado y, perteneciente a la clase T (clase objetivo).
- Buscamos aprender una función f: D -> T (hipótesis) que minimice la discrepancia entre sus predicciones y los valores de y del conjunto de datos
- Cuando T es discreta, hablamos de **clasificación**. Si T es continua (en general, los números reales), es un problema de **regresión**.

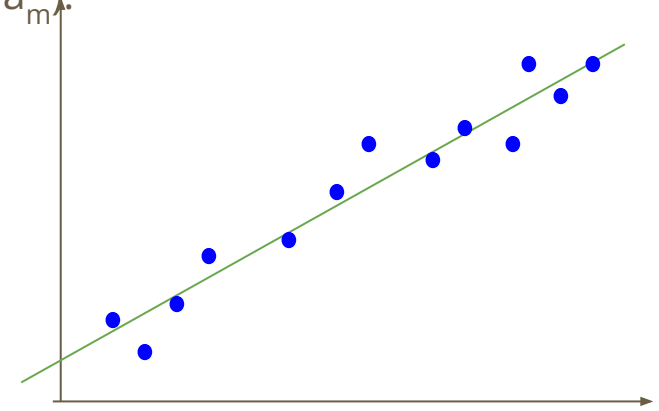
# Ejemplos

- ¿Cuál es el valor de venta de una casa de 2 dormitorios, en Atlántida?
- ¿Cuál va a ser la velocidad del viento el próximo domingo a las 17 h en el puerto?
- ¿Cuántas personas van a pasar por la terminal de ómnibus de Tres Cruces el martes?
- ¿Cuánto va a ser el consumo de energía eléctrica en los hogares de Uruguay la próxima semana?
- ¿Cuánto va a ser el valor del dólar el miércoles?

# Regresión lineal

- Es el modelo más simple que se puede construir a partir de datos.
- Se busca un modelo de la siguiente forma: $y = a_m\ x_m + a_{m-1}\ x_{m-1} + \ldots + a_0 = XA$

y: variable de salida, $x_i$: variables de entrada, $a_i$ coeficientes. X: (1, $x_1$, $x_2$, ..., $x_m$), A: ($a_0$, $a_1$, ..., $a_m$).



1 variable de entrada: RL simple

Más de 1 variable de entrada: RL multivariada

# Regresión lineal usando mínimos cuadrados (OLS)

Se quiere hallar la línea que minimice el error cuadrático medio (MSE)

n: cantidad de datos

$$MSE = \frac{1}{2n}\sum_{i=1}^{i=n}\left(y_{data}^{i} - y_{pred}^{i}\right)^2$$

Es lo mismo que minimizar la suma de los errores al cuadrado (SSE)

$$SSE = \sum_{i=1}^{i=n}\left(y_{data}^{i} - y_{pred}^{i}\right)^2$$

# Mínimos cuadrados

Se puede reescribir la ecuación anterior en función del modelo:

$$SSE = \sum_{i=1}^{i=n} \left(y_{data}^{i} - X_{data}^{i} A\right)^2$$

Reescribiendo la sumatoria en forma vectorial (matricial):

$$SSE = (y_{data} - X_{data} A)^T (y_{data} - X_{data} A)$$

# Mínimos cuadrados

Para buscar el mínimo de una función buscamos en donde la derivada se hace 0:

$$\frac{\partial SSE}{\partial A} = -2\, X_{data}^{T}(y_{data} - X_{data}A)$$

$$0 = -2\, X_{data}^{T}\left(y_{data} - X_{data}\hat{A}\right)$$

Despejando:

$$\hat{A} = \left(X_{data}^{T}X_{data}\right)^{-1}X_{data}^{T}y_{data}$$

# Mínimos cuadrados

Existen varios solvers para resolver este problema:

El más común es OLS → sklearn.regression.linearmodel

Esto sucede porque la inversión de la matriz es muy costosa computacionalmente, y en casos donde se tiene grandes volúmenes de datos, no es posible hallar los coeficientes de forma matricial, sino que con métodos iterativos.

$$\hat{A} = \left( X_{data}^{T} X_{data} \right)^{-1}$$

En sklearn existen varios: ridge, RANSAC, etc...

# Regresión lineal - métodos

Es bueno en presencia de outliers

- RANSAC: is an iterative algorithm for the robust estimation of parameters from a subset of inliers from the complete data set.(sklearn ransac)
- Ridge: This model solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm. Also known as Ridge Regression or Tikhonov regularization. This estimator has built-in support for multi-variate regression.(sklearn ridge)
- Lasso: Linear Model trained with L1 prior as regularizer (aka the Lasso).(sklearn lasso)

Intentan penalizar el overfitting, intentando no generar modelos complejos. Lasso se utiliza más como feature selection

# $R^2$ - Coeficiente de determinación

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{i=n}\left(y_{data}^{i} - y_{pred}^{i}\right)^2}{\sum_{i=1}^{i=n}\left(y_{data}^{i} - \overline{y_{data}}\right)^2} = 1 - \frac{MSE}{Var(y_{data})}$$

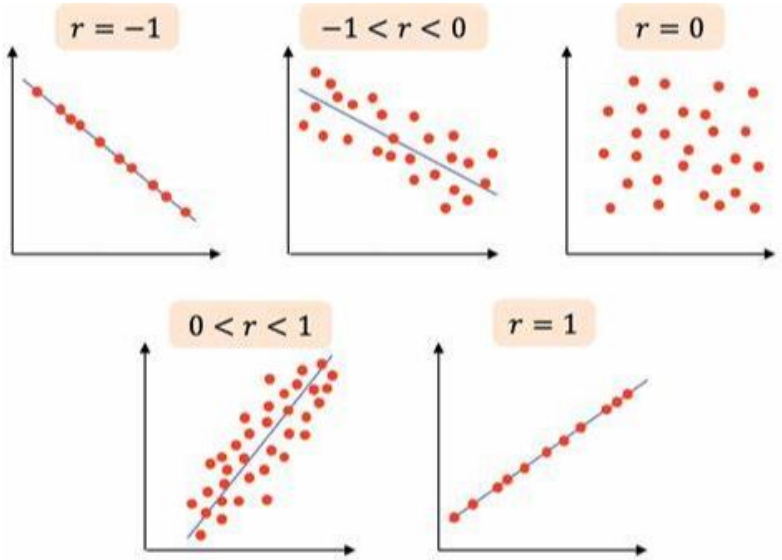
- En su concepto más simple es una medida del error (Cuando MSE=0, $R^2$=1).
- En una regresión lineal tiene un significado estadístico.
- Es la fracción de la varianza total de la respuesta que puede ser capturada por el modelo.

# Prueba t de Student

- La prueba t de *Student* considera como hipótesis nula ($H_0$) que el coeficiente estimado por la regresión lineal de un término i ($a_i$) es 0, y como hipótesis alterna que el coeficiente $a_i$ es distinto a 0, i varía entre 1 y m.
- Se calcula el valor p, se utiliza un valor de significancia ($\alpha$, usualmente 0,05). Si $p \geqslant \alpha$, se acepta $H_0$, y se concluye que $a_i$ es 0 y el término considerado se puede descartar.

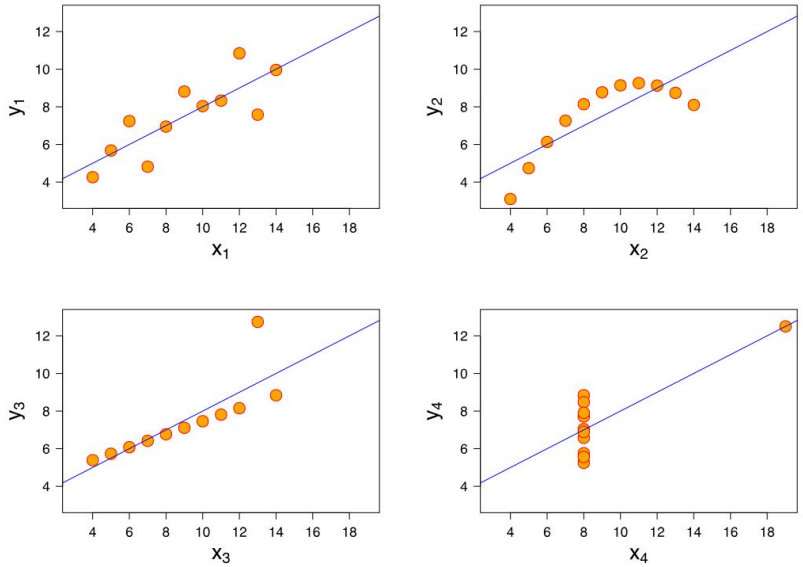
# Correlaciones

- Pearson: es una medida estadística que evalúa la **relación lineal entre dos variables continuas**. Su valor oscila entre -1 y 1, donde 1 indica una correlación positiva perfecta, -1 una correlación negativa perfecta y 0 indica que no hay correlación. Para calcularlo, se utiliza la fórmula que relaciona las covarianzas de las variables con sus desviaciones estándar.



# Ejemplo del Dinosaurio

- Anscombe's quartet - Wikipedia
- Datasaurus dozen - Wikipedia, Artículo, Imágenes



# La próxima clase ...

## Regresión logística