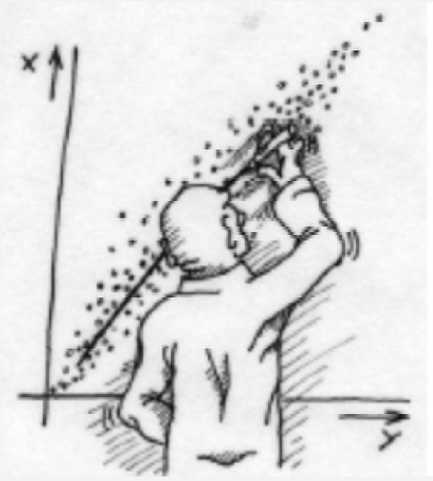
Predicción Lineal



La predicción lineal es una operación matemática donde los valores futuros de una señal de tiempo discreto se estiman como una función lineal de muestras anteriores.

Modelamiento Predictivo

$$(X_1, Y_1), \cdots, (X_n, Y_n) \in \Xi \times Y$$

- Usualmente $\Xi = \mathbb{R}^p$
- $Y = \mathbb{R}$ en este caso hablamos de regresión
- $Y = \{0,1\}$ o $Y = \{-1,1\}$ en este caso hablamos de clasificación binaria.
- $Y = \{1,2,\cdots,m\}$ en este caso es clasificación multiclase

Predicción Lineal

Regresión

Y puede ser el precio de una vivienda, X características de la casa y del entorno: el número de dormitorios, el tamaño de la casa, tasa de delincuencia del barrio, etc.

Clasificación

Los valores Y son 'etiquetas' que indican al grupo al que pertenece X. Los valores de Y podrían ser las 'etiquetas' que podemos asignar a un cliente bancario {fraude, no fraude}, y las X son características observables sobre los clientes: edad, profesión, etc.

METRICAS DE CALIDAD PREDICTIVAS

El error cuadrático medio (RMSE) mide la cantidad de error que hay entre dos conjuntos de datos.
Error porcentual medio absoluto (MAPE), es el promedio del error absoluto o diferencia entre la demanda real y el pronóstico.

METRICAS DE CALIDAD PREDICTIVAS (cont)

El error absoluto medio (MAE) es una medida de la diferencia entre dos variables continuas.
La validación cruzada generalizada (GCV), es una técnica que consiste en dividir los datos en varios conjuntos de datos y luego elegir uno de los conjuntos para medir la precisión de la predicción "test" y el resto para entrenar.

3.3 Estadísticos de Asociación

Covarianza

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

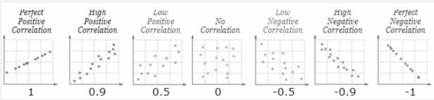
Es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.

Correlación

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

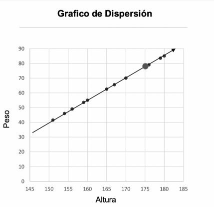
Indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas

Predicción Lineal | Grados de Correlación



La función de correlación entrega valores comprendidos entre -1 y 1, mientras que la covarianza entregará valores coherentes con la escala de la multiplicación de las variables.

Predicción Lineal | Ejemplo



¿Cuál sería el valor del Peso con una altura de 175 cm?

Peso = 1.5 * Altura - 185

Peso = 1.5 * 175 – 185

77.5 = 1.5 * 175 - 185

Por lo tanto, una persona que mide 175 cm, tendrá un peso de 77.5 kg

Modelo de Regresión

Un **modelo de regresión** es un modelo que permite describir cómo influye una variable X sobre otra variable Y.

El objetivo es obtener predicciones razonables de Y para distintos valores de X a partir de una muestra de n pares de valores $(x_1,y_1), \ldots, (x_n,y_n)$.

Predicción Lineal Múltiple

Modelo de regresión lineal múltiple extiende el modelo de regresión lineal simple con la incorporación de dos o más covariables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i$$

- β_0 es el término constante
- β_1 a β_k son k parámetros
- e es el término de error
- El supuesto de media condicional cero se mantiene: $E(e_i | X_{i1}, X_{i2}, \ldots, X_{ik}) = 0$
- Igual que antes, minimizamos la suma de residuales cuadrados, de modo que tenemos $p=k+1$ condiciones de primer orden (o $p=k+1$ parámetros a estimar)

Estimación de Parámetros por MCO

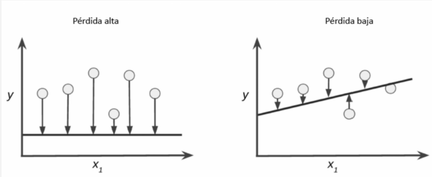
Función de pérdida

Entrenar un modelo simplemente significa aprender (determinar) valores correctos para todas los parámetros β_0 y β_1 para los datos de la muestra.

La pérdida es una penalidad por una predicción incorrecta. Esto quiere decir que la pérdida es un número que indica qué tan incorrecta fue la predicción del modelo en un solo par de valores (x_i,y_i).

El objetivo de entrenar un modelo es encontrar β_0 y β_1 que, en promedio, tengan pérdidas bajas en todos los datos $D=\{(x_i,y_i): i=1, \ldots, n\}$

Función Pérdida



Métricas de Calidad Predictivas

Suma de cuadrados

Podemos descomponer cada observación como la suma de dos términos: un componente explicado y un componente no explicado, es decir:

$$Y_i = \hat{Y}_i + \left(Y_i - \hat{Y}_i\right) = \hat{Y}_i + e_i.$$

De modo que podemos definir lo siguiente:
- $\sum_{i=1}^n (Y_i - \bar{Y})^2$ es la suma de cuadrados totales (SCT)
- $\sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2$ es la suma de cuadrados explicada (SCE)
- $\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2$ es la suma de cuadrados residual (SCR)

Lo anterior implica que SCT=SCE+SCR.

Nota: SCT es la suma de "desviaciones al cuadrado" de las observaciones de la muestra: es proporcional, más no igual, a VAR(Y).