Muestreo

La manera más directa de reducir el tamaño de una población o conjuntos de individuos es realizar una selección o muestreo.

Nos podemos plantear dos situaciones, dependiendo de la disponibilidad de la población:

- Se dispone de la población: en este caso se ha de determinar qué cantidad de datos son necesarios y cómo hacer la muestra.

- Los datos son ya una muestra de realidad y sólo representan una parte de esta realidad.

Muestreo

- Muestreo Aleatorio Simple: Cualquier instancia tiene la misma probabilidad de ser extraída en la muestra. Dos versiones, con reemplazamiento y sin reemplazamiento.

- Muestreo Aleatorio Estratificado: El objetivo de este muestreo es obtener una muestra balanceada con suficientes elementos de todos los estratos, o grupos. Una versión simple es realizar un muestreo aleatorio simple sin reemplazamiento de cada estrato hasta obtener los *n* elementos de ese estrato. Si no hay suficientes elementos en un estrato podemos utilizar en estos casos muestreo aleatorio simple con reemplazamiento (sobremuestreo).

Muestreo

- Muestreo de Grupos: El muestreo de grupos consiste en elegir sólo elementos de unos grupos. El objetivo es generalmente descartar ciertos grupos que, por diversas razones, pueden impedir la obtención de buenos modelos.

- Muestreo Exhaustivo: Para los atributos numéricos (normalizados) se genera al azar un valor en el intervalo posible; para los atributos nominales se genera al azar un valor entre los posibles. Con esto obtenemos una instancia ficticia y buscamos la instancia real más similar a la ficticia. Se repite este proceso hasta tener *n* instancias. El objetivo es cubrir completamente el espacio de instancias.

Muestreo

- Cuántos datos son necesarios mantener?

Depende, en general, del número de “grados de libertad” (número de atributos y valores) y del método de aprendizaje y de su expresividad (por ejemplo una regresión lineal requiere muchos menos ejemplos que una red neuronal).

Se utiliza una estrategia incremental, en el que se va haciendo la muestra cada vez más grande (y diferente si es posible) hasta que se vea que los resultados no varían significativamente entre un modelo y otro.

Minería de Datos

Dificultades

Aparte del gran volumen, ¿por qué algunas técnicas de aprendizaje automático y estadísticas no son directamente aplicables?

- Los datos residen en el disco. No se pueden escanear múltiples veces.
- Algunas técnicas de muestreo no son compatibles con algoritmos no incrementales.
- Muy alta dimensionalidad.
- DATOS IMPERFECTOS...

Datos Imperfectos

Ruido

- En la evidencia o ejemplos de entrenamiento.
- Erróneos valores de argumentos de los ejemplos.
- Clasificación errónea de algún ejemplo.
- En el conocimiento previo.

Ejemplos de entrenamiento muy dispersos.

Conocimiento previo correcto pero inapropiado.

- Existencia de muchos predicados irrelevantes para el problema.
- Conocimiento previo insuficiente para el problema a aprender (algunos predicados auxiliares serían necesarios).

Argumentos faltantes (nulos) en los ejemplos.

Tipos de conocimiento

•Asociaciones: Una asociación entre dos atributos ocurre cuando la frecuencia con la que se dan dos valores determinados de cada uno conjuntamente es relativamente alta.

Ejemplo: en un supermercado se analiza si los pañales y juguetes de bebé se compran conjuntamente.

•Dependencias: Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Ojo! Existen muchas dependencias nada interesantes (ojo con causalidades inversas).

Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.

La búsqueda de asociaciones y dependencias se conoce como análisis exploratorio.

Tipos de conocimiento

•Clasificación: Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas.

Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, número de dioptrías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria.

Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.

•Segmentación: La segmentación (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

Tipos de conocimiento

•Tendencias: El objetivo es predecir los valores de una variable continua a partir de la evolución de otra variable continua, generalmente el tiempo.

Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.

•Información del Esquema: (descubrir claves primarias alternativas, R.I.).

•Reglas Generales: patrones que no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

Regresión

Modelo Estadístico

x_i variables explicativas, y_i variable respuesta:

$$y_i = r(x_{il}, \ldots, x_{ip}) + \varepsilon_i$$

- r es la función determinista que explica el comportamiento de y en función de las variables explicativas x_i
- ε_i es la aleatoriedad en el comportamiento de cada individuo (o error)

Modelo Estadístico

El error para cada individuo se caracteriza a partir de una distribución de probabilidad generadora de los distintos valores y cuyo valor esperado es 0:

$$E[\varepsilon_i] = E[y_i - r(x_{il}, \ldots, x_{ip})] = 0$$

Buscamos una función tal que, en promedio, las desviaciones al cuadrado respecto de los puntos sean mínimas.

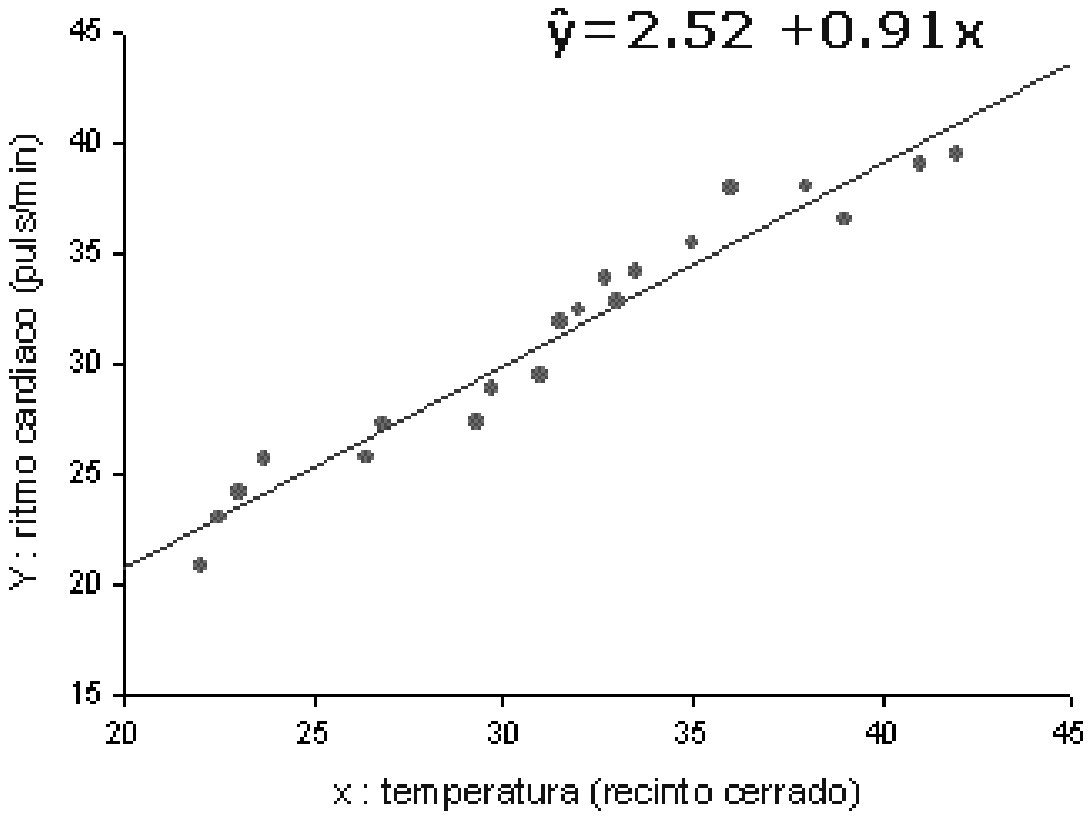
Los modelos difieren dependiendo de la variable respuesta: numérica, binaria o categórica.

Modelo de Regresión

Modelo de regresión es aquel en el que las variables explicativas y la variable respuesta son todas cuantitativas.

$$y_i = r(x_{il}, \ldots, x_{ip}) + \varepsilon_i$$

- r es la función determinista que explica el comportamiento de y en función de las variables explicativas x_i
- ε_i es la aleatoriedad en el comportamiento de cada individuo (o error)


ŷ = 2.52 + 0.91x
Y : ritmo cardiaco (puls/min)
x : temperatura (recinto cerrado)
45
40
35
30
25
20
15
20
25
30
35
40
45

Modelo de Regresión

El error para cada individuo:

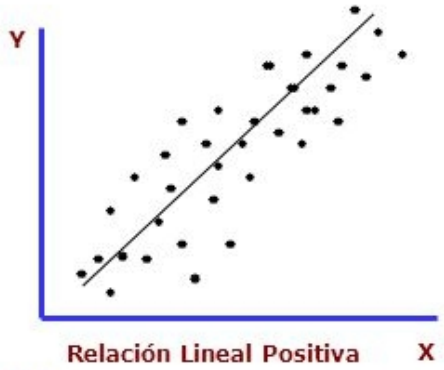
$$E[\varepsilon_i] = E[y_i - r(x_{il}, \ldots, x_{ip})]$$

Buscamos una función tal que, en promedio, las desviaciones al cuadrado respecto de los puntos sean mínimas.

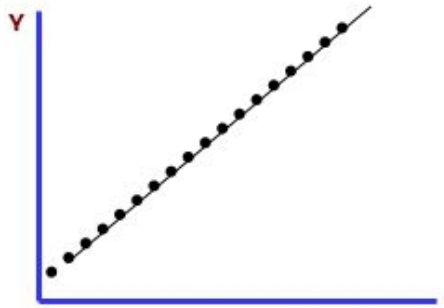
Minimizamos el error cuadrático medio.

La función de regresión: $min_r\ E[(y_i - r(x_{il}, \ldots, x_{ip}))^2]$

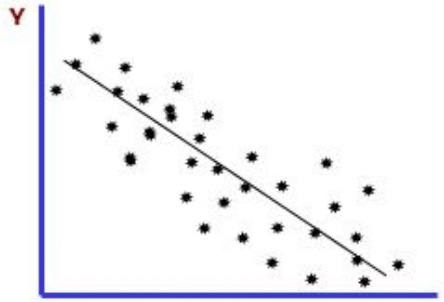
Relación entre variables cuantitativas



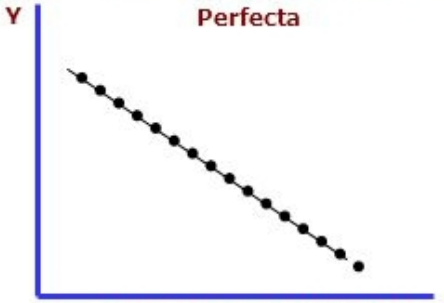
Relación Lineal Positiva



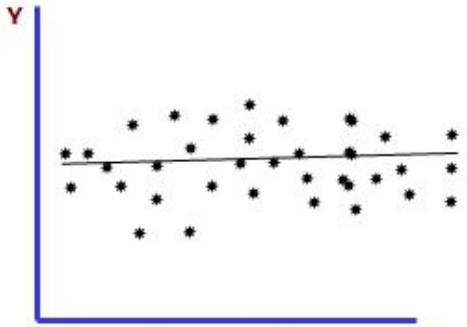
Relación Lineal Positiva Perfecta



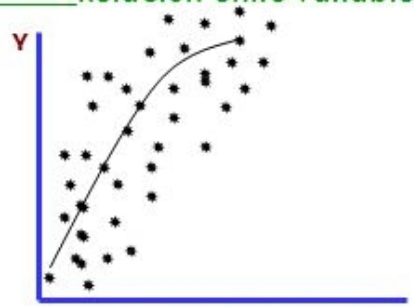
Relación Lineal Negativa



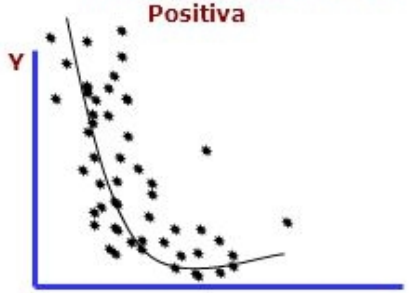
Relación Lineal Negativa Perfecta



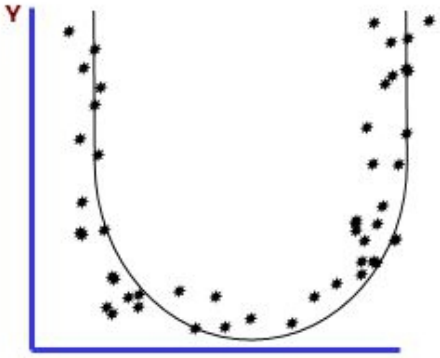
No exixte relación



Relación Curvilinea Positiva



Relación Curvilinea Negativa



Relación Curvilinea Positiva

Simple y Múltiple

Regresión simple: una sola variable explicativa

Regresión múltiple: varias variables explicativas

Regresión Lineal

Cada variable explicativa participa de forma aditiva y constante para todo el dominio.

Regresión múltiple: varias variables explicativas

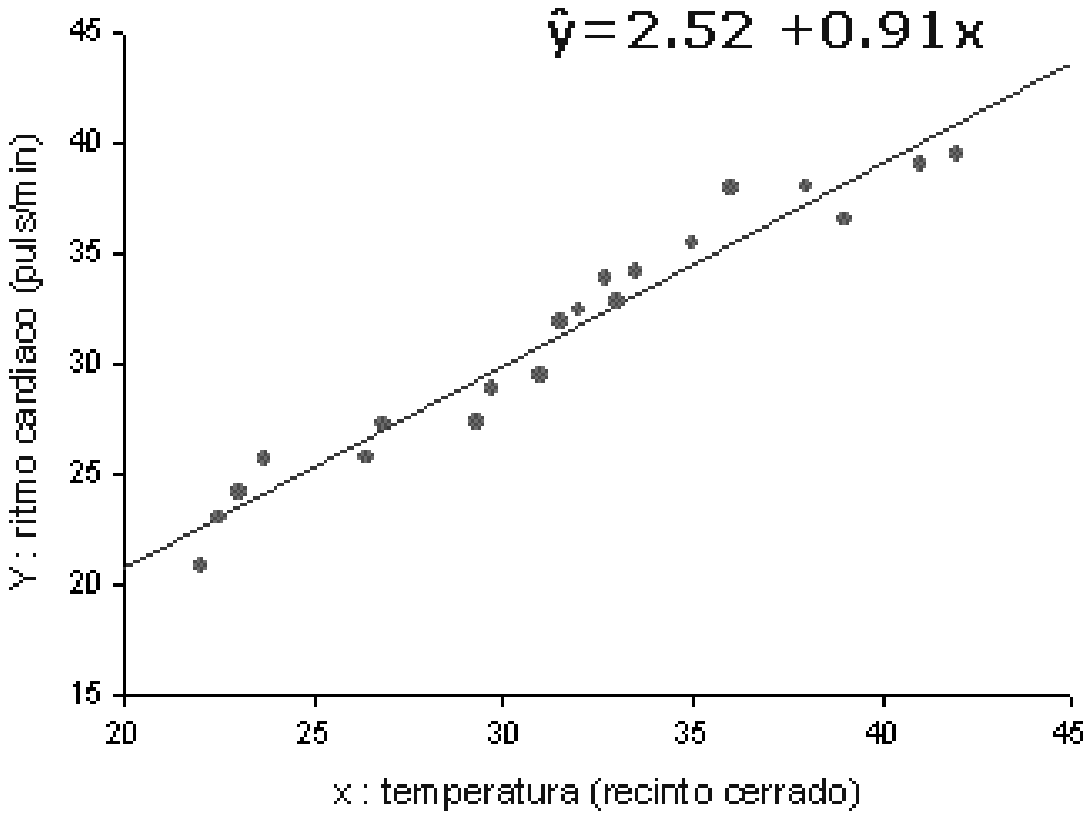
Simple pero la variable explicativa pueden ser transformaciones de las originales: logarítmicas, raíz cuadrada, polinómicas, etc.

Se puede tomar variables explicativas discretizadas (binning).

Evaluación

Varianza residual s^2, valores pequeños indican modelos precisos alrededor de los datos muestrales.

Descomposición de la suma de cuadrados de la variable de respuesta que se presenta en forma de tabla como “Análisis de Varianza” o AnoVa.


ŷ = 2.52 + 0.91x
Y : ritmo cardiaco (puls/min)
x : temperatura (recinto cerrado)
15
20
25
30
35
40
45
20
25
30
35
40
45