DIPLOMADO EN BIG DATA PARA LA TOMA DE DECISIONES

Introducción al Modelamiento Estadístico

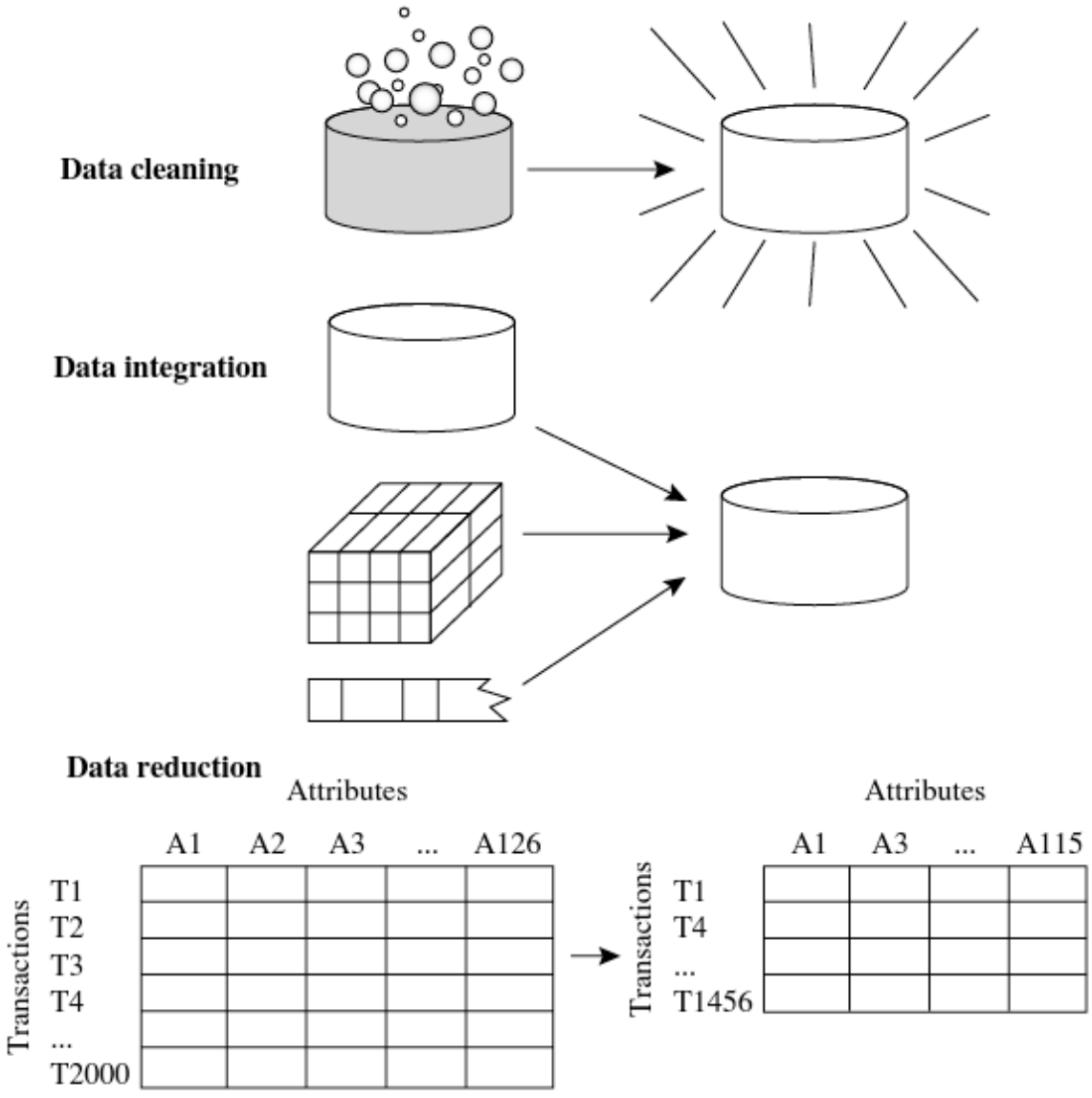
Parte 1: Preprocesamiento de Datos

1.1.- Introducción

Introducción

- Las bases del mundo real están susceptibles a ruido, valores faltantes y datos inconsistentes.
- Razones: gran tamaño y múltiples orígenes.
- Datos de mala calidad $\Longrightarrow$ conclusiones de mala calidad.
- Hay varias técnicas de preprocesamiento:
  - limpieza de datos
  - integración de datos
  - reducción de datos
  - transformación de datos

Introducción



Introducción: Calidad de los Datos

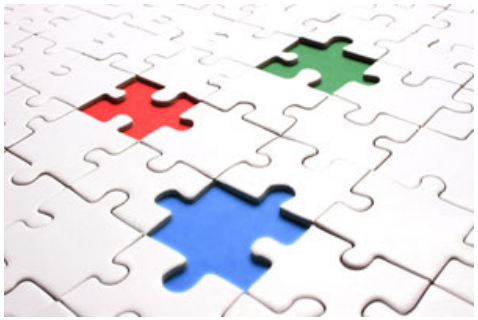
- Precisión: instrumentos, errores humanos o computacionales, ingreso consiente de información incorrecta (datos faltantes disfrazados), limitaciones tecnológicas.
- Completitud: falta de información para algunos atributos, información no recolectada a tiempo, errores de equipamiento, malinterpretaciones.
- Consistencia: inconsistencias de formatos (fechas), variables nombradas diferentes, registros con diferentes criterios.
- Pertinencia: disposición de la información fuera de plazo.
- Credibilidad: múltiples errores pasados.
- Interpretabilidad: códigos poco claros.

1.2.- Limpieza de Datos

1.2.1.- Datos Faltantes

Datos Faltantes

- Puede ser un problema muy serio (aunque sean muy pocos los casos) y la viabilidad de su solución depende del tipo de mecanismo de generación de datos faltantes.



- Soluciones posibles:
  - Eliminación de unidades o filas de la matriz de datos con datos faltantes en algunas variables.
  - Uso de métodos de imputación de datos faltantes.
  - Uso de modelos estadísticos que hacen uso de toda la información disponible.

Datos Faltantes: Mecanismos

- **Missing Completely at Random (MCAR):** Ocurre cuando los eventos que llevan a la falta del dato en la variable o atributo de una unidad son independientes de los datos observados y no observados en el conjunto de datos, y ocurren enteramente al azar $\Longrightarrow$ **todos los procedimientos anteriores son válidos (aunque no necesariamente eficientes)**.

- **Missing Random (MAR):** Ocurre cuando los eventos que llevan a la falta del dato en la variable o atributo de una unidad no son aleatorios, pero pueden explicarse en forma completa por los datos efectivamente observados $\Longrightarrow$ **sólo algunos de los procedimientos anteriores son válidos.**

Datos Faltantes: Mecanismos

- **Missing Not at Random (MNAR):** Ocurre cuando los eventos que llevan a la falta del dato en la variable o atributo de una unidad no son aleatorios y dependen de lo que no es observado $\Longrightarrow$ **ninguno de los procedimientos existente son válidos, sólo se pueden hacer estudios de sensibilidad!!**

- El problema: No es posible verificar el mecanismo de generación de datos faltantes en base a los datos disponibles. Se requiere de información externa!!!

1.2.2.- Datos Ruidosos

Datos ruidosos

- **“Ruido”:** es un error aleatorio o varianza en una variable.
- Herramientas gráficas como boxplot o gráficos de dispersión ayudan en la detección de datos atípicos (outliers), que pueden representar ruido.

¿Cómo podemos suavizar los datos para remover el ruido?

Técnicas de Suavizamiento: Binning

- `Binning:` suavizan datos ordenados con información de sus vecinos.
- Los valores ordenados son distribuidos en “cajones” o `bins`.
- Realizan suavizamiento local.

Ejemplo:

Se tiene información de los precios (US$) de los artículos vendidos en una tienda:

4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

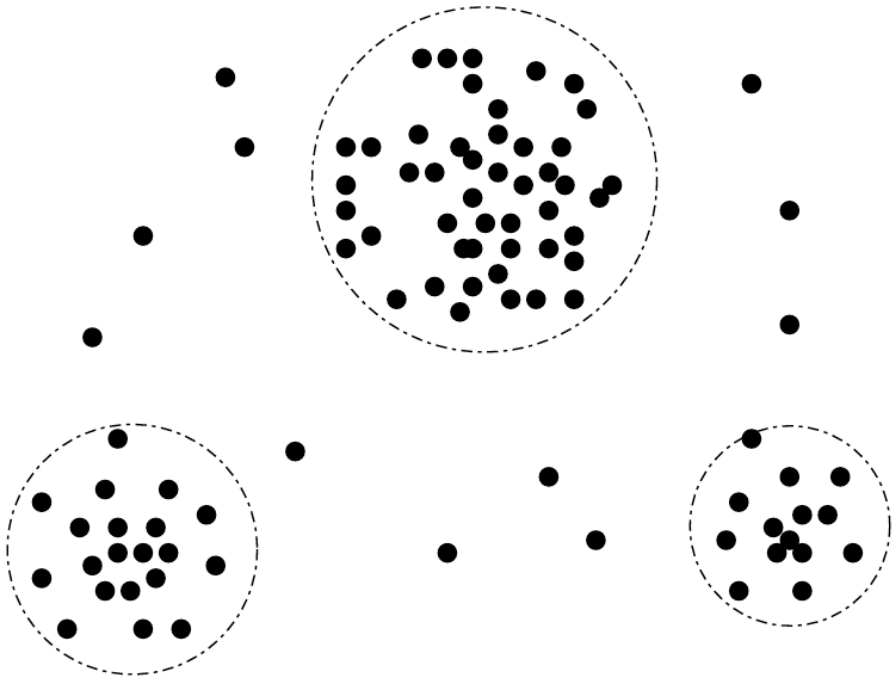
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Técnicas de Suavizamiento: Regresión

- El suavizamiento también puede hacerse mediante regresión.
- Esta técnica ajusta los datos a una función.
- Regresión lineal busca encontrar la “mejor” línea que ajuste un par de variables, de manera de uno de ellos sirva para predecir el valor del otro.
- Una extensión es la regresión múltiple, donde más de dos variables se consideran para ajustar una superficie multidimensional.

Técnicas de Suavizamiento: Detección de Outliers

- Los outliers pueden detectarse con técnicas de agrupamiento.



1.3.- Integración de Datos

Integración de Datos

- Usualmente se requiere la integración de datos: fundir datos desde diferentes fuentes.
- Reducir y evitar redundancias e inconsistencias en la base de datos final.
- Ganancia de tiempo.

Integración de Datos: Problema de identificación de la entidad

- Los datos provienen de diferentes fuentes de origen.
- La integración de esquemas y coincidencia de objetos puede ser compleja.
- Este proceso también está íntimamente relacionado con la limpieza de datos.
- Especial cuidado con los formatos y definiciones de las variables a fundir.

Integración de Datos: Problema de identificación de la entidad

Ejemplos:

1. `id_cliente` $\Longleftrightarrow$ `num_cliente`
2. `tipo_pago=H y S` $\Longleftrightarrow$ `tipo_pago=1 y 2`

Integración de Datos: Redundancia y Correlación

- Una variable es redundante si ésta puede derivarse de otras variables presentes en la base de datos.
- Inconsistencias en las variables o nombres también producen redundancia.
- Algunas redundancias pueden detectarse mediante análisis de correl es decir, en base a los datos disponibles, se puede medir qué tan fuerte es la asociación lineal entre dos variables.
- Datos nominales: test de χ^2.
- Datos numéricos: coeficiente de correlación y covarianza.

1.4.- Reducción de Datos

Reducción de Datos

- Cuando se recoleta la información, lo más frecuente es tomar el mayor número posible de variables.
- Al tomar demasiadas variables, usualmente ocurre que estén correlacionadas o midan lo mismo bajo diferentes puntos de vista.
- Las técnicas de reducción de datos buscan obtener una representación reducida de los datos, manteniendo la integridad del grupo completo.
- La reducción de dimensionalidad es el proceso de reducir el número de variables aleatorias a considerar.
- Uno de los métodos más populares, es el análisis de componentes principales.

1.4.1.- Componentes Principales

Componentes Principales

- Fue propuesto originalmente por Pearson (1901) y desarollado posteriormente por Hotelling (1933, 1936).
- Se aplica cuando suponemos que no existe causalidad entre componentes de observaciones multivariadas.

Componentes Principales

- Idea principal: transformar el conjunto original de variables en otro conjunto de nuevas variables no correlacionadas entre sí.
- Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.
- Se buscan $m < p$ variables que sean combinaciones lineales de las p originales, que no estén correlacionadas y que recojan la mayor parte de la información o variabilidad de los datos.

Componentes Principales

Supongamos que estamos interesados en conocer la estructura de dependencia conjunta de las componentes de la variable $p-$variada

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^t,$$

con

$$E(\boldsymbol{X}) = \boldsymbol{\mu}, \qquad var(\boldsymbol{X}) = \boldsymbol{\Sigma},$$

donde $r(\boldsymbol{\Sigma}) = p$ y sus valores propios,

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p,$$

son todos diferentes.

Componentes Principales

La información sobre la estructura de dependencia de las variables se encuentra contenida en las matriz $\mathbf{\Sigma}$.

También se podría trabajar con la matriz de correlaciones.

Dado que éstas son desconocidas, basaremos nuestra inferencia en sus estimadores insesgados.

Componentes Principales

Supongamos que disponemos de una muestra de tamaño N,

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \vdots \\ X_{N1} & \dots & X_{Np} \end{pmatrix}.$$

El estimador insesgado de $\boldsymbol{\Sigma}$ corresponde a

$$\boldsymbol{S} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^t.$$

Componentes Principales

Definición: La primera componente principal de la matriz $\boldsymbol{S}$ corresponde a la combinación lineal,

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ &= \boldsymbol{a}_1^t \boldsymbol{X}, \end{aligned}$$

cuya varianza muestral,

$$\begin{aligned} S_{Y_1}^2 &= \sum_{i=1}^{p}\sum_{j=1}^{p} a_{1i}a_{1j}s_{ij} \\ &= \boldsymbol{a}_1^t \boldsymbol{S}\boldsymbol{a}_1 \end{aligned}$$

es la mayor entre las varianzas de todas las combinaciones lineales posibles, sujeto a

$$\boldsymbol{a}_1^t \boldsymbol{a}_1 = 1.$$

Componentes Principales

Debemos maximizar la función

$$S_{Y_1}^2 + l_1(1 - \boldsymbol{a}_1^t \boldsymbol{a}_1).$$

Derivando e igualando a cero, obtenemos,

$$2(S - l_1 \boldsymbol{I})\boldsymbol{a}_1 \quad = \quad 0$$

Por el teorema de Roché-Frobenius, para que este sistema tenga solución $\neq 0$, la matriz $(S - l_1 \boldsymbol{I})$ debe ser singular, lo que implica que:

$$|S - l_1 \boldsymbol{I}| = 0$$

Componentes Principales

Luego, l_1 es un valor propio de $\boldsymbol{S}$ y $\boldsymbol{a}_1$ su vector propio asociado. Premultiplicando por $\boldsymbol{a}_1^t$,

$$\begin{aligned} \boldsymbol{a}_1^t(S - l_1\boldsymbol{I})\boldsymbol{a}_1 &= 0 \\ \Longleftrightarrow \qquad S_{Y_1}^2 &= l_1, \end{aligned}$$

luego, para mximizar la varianza, l_1 debe ser el mayor valor propio de $\boldsymbol{S}$.

Componentes Principales

Luego, la **primera componente principal** de la matriz $\boldsymbol{S}$ corresponde a la combinación lineal

$$Y_1 = \boldsymbol{a}_1^t \boldsymbol{X},$$

donde $\boldsymbol{a}_1$ corresponde al vector propio de $\boldsymbol{S}$ asociado a su mayor valor propio. Impondremos que $\boldsymbol{a}_1^t \boldsymbol{a}_1 = 1$.

La varianza muestral de Y_1 corresponde a l_1.

Componentes Principales

La segunda componente principal corresponde a la combinación lineal

$$Y_2 = \boldsymbol{a}_2^t \boldsymbol{X},$$

sujeto a

$$\boldsymbol{a}_2^t \boldsymbol{a}_2 = 1, \qquad \boldsymbol{a}_2^t \boldsymbol{a}_1 = 0,$$

donde $\boldsymbol{a}_2$ se elige de modo que maximize su varianza muestral,

$$S_{Y_2}^2 = \boldsymbol{a}_2^t \boldsymbol{S} \boldsymbol{a}_2.$$

Obtendremos que $\boldsymbol{a}_2$ debe elegirse como el vector propio asociado al segundo valor propio de la matriz $\boldsymbol{S}$, al ordenarlos en orden decreciente de magnitud.

Componentes Principales

Definición: La *j*-ésima componente principal de la muestra de observaciones p-variadas $\boldsymbol{X}$, se define como la combinación lineal

$$Y_j = \boldsymbol{a}_j^t \boldsymbol{X}$$

donde el vector $\boldsymbol{a}_j$ corresponde al vector propio de la varianza muestral $\boldsymbol{S}$, correspondiente al j-ésimo valor propio de $\boldsymbol{S}$, al ordenarlos en orden decreciente de magnitud.

Se verifica que la varianza muestral de Y_j corresponde a

$$S_{Y_j}^2 = l_j,$$

su valor propio asociado.

Componentes Principales

Se define la varianza total como

$$tr(\boldsymbol{S}) = \sum_{j=1}^{p} l_j.$$

La importancia de la j-ésima componente se mide como

$$\frac{l_j}{tr(\boldsymbol{S})}.$$

La magnitud y signo de los coeficientes asociados nos entregan información sobre la importancia de cada variable y su signo el sentido.

Se verifica que

$$\boldsymbol{S} = l_1 \boldsymbol{a}_1 \boldsymbol{a}_1^t + \ldots + l_p \boldsymbol{a}_p \boldsymbol{a}_p^t.$$

Componentes Principales

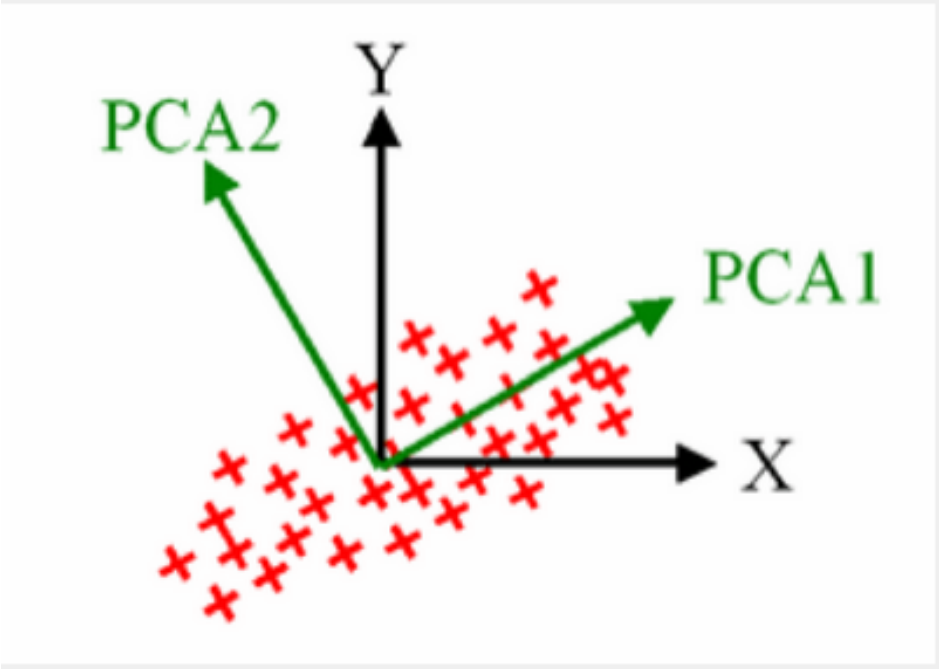
- Si se trabaja con la matriz de correlaciones, todos los elementos de la diagonal son iguales a 1.
- Esto implica que $tr(\boldsymbol{S}) = p$ y la proporción de varianza explicada por la j-ésima componente es $\frac{l_j}{p}$.

Componentes Principales

¿Para qué sirve todo esto?

- Hemos encontrado una base de variables ortogonales entre sí, que son ordenables de acuerdo a la medida en que explican la variabilidad de la muestra, y que en conjunto explican la totalidad de esta variablidad.
- Esto nos permite interpretar el problema en términos de variables independientes (ortogonales).
- De acuerdo al % de varianza muestral que deseemos explicar, podemos retener sólo $r \leq p$ componentes y con esto, disminuir la dimensionalidad del problema.

Componentes Principales



Componentes Principales: Ejemplo

El heptatlón es una prueba olímpica para mujeres que involucra 7 disciplinas:

- vallas
- salto alto
- tiro de la bala
- 200 m planos
- salto largo
- lanzamiento de la jabalina
- 800 m planos

Componentes Principales: Ejemplo

Los datos que tenemos provienen de los resultados de esta prueba en los juegos olímplicos de Seúl 1988, cuya ganadora fue Jackie Joyner-Kersee.

Componentes Principales: Ejemplo

- Estamos interesado en la relación entre los resultados de las diferentes pruebas del heptatlón.
- ¿Habrá algún factor “velocidad” o de “salto” o de “tiro” en el puntaje total?

Componentes Principales: Ejemplo

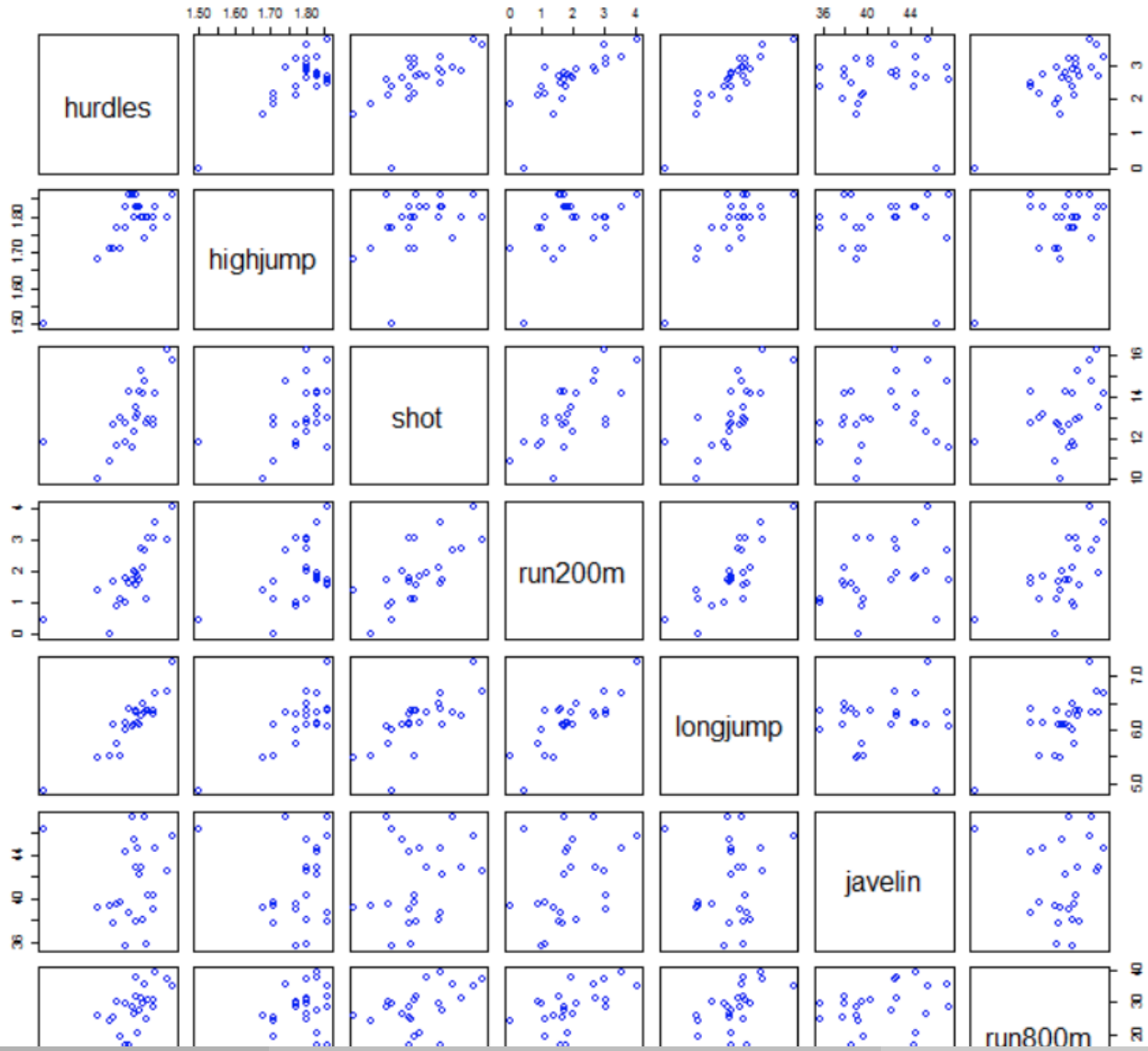
La matriz de correlación entre las variables es:

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.00	0.81	0.65	0.77	0.91	0.01	0.78
highjump	0.81	1.00	0.44	0.49	0.78	0.00	0.59
shot	0.65	0.44	1.00	0.68	0.74	0.27	0.42
run200m	0.77	0.49	0.68	1.00	0.82	0.33	0.62
longjump	0.91	0.78	0.74	0.82	1.00	0.07	0.70
javelin	0.01	0.00	0.27	0.33	0.07	1.00	-0.02
run800m	0.78	0.59	0.42	0.62	0.70	-0.02	1.00

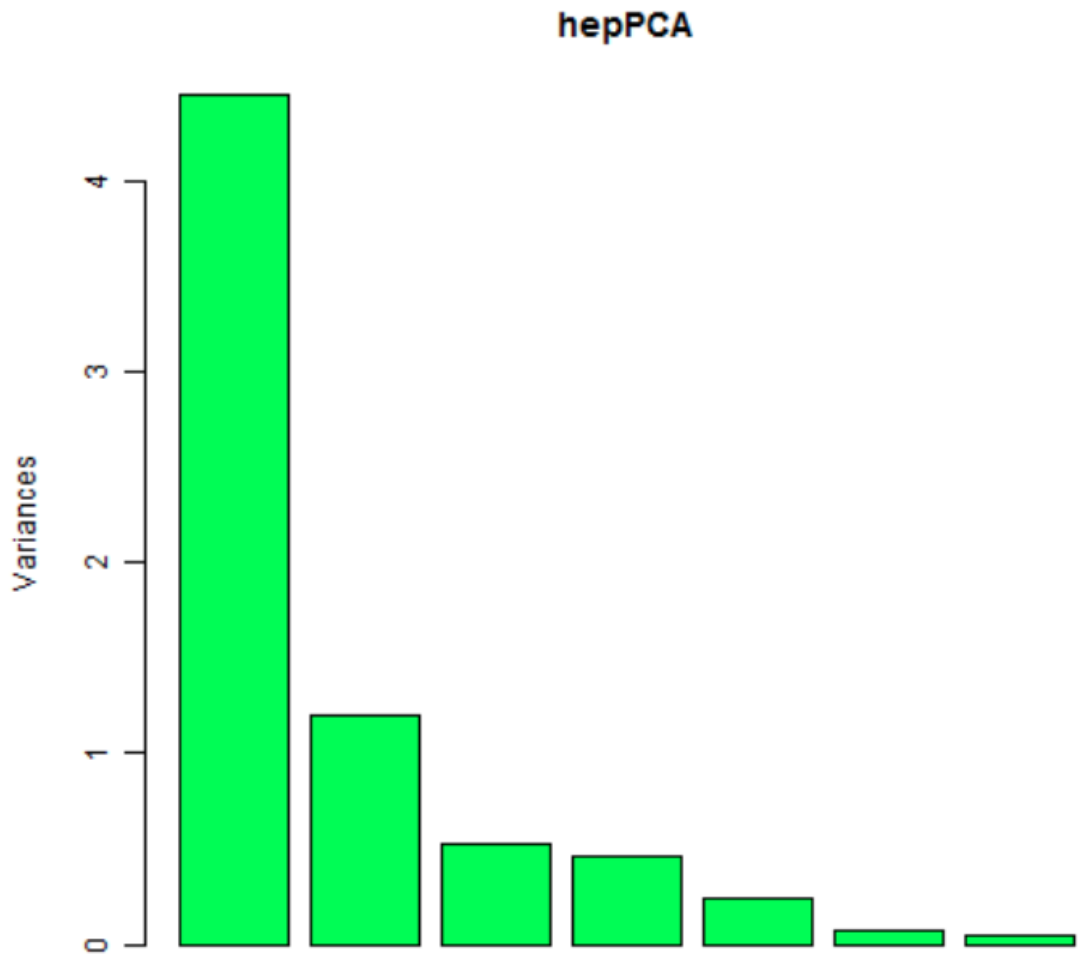
Componentes Principales: Ejemplo

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles		0.81	0.65	0.77	0.91		0.78
highjump	0.81		0.44	0.49	0.78		0.59
shot	0.65	0.44		0.68	0.74	0.27	0.42
run200m	0.77	0.49	0.68		0.82	0.33	0.62
longjump	0.91	0.78	0.74	0.82			0.70
javelin			0.27	0.33			
run800m	0.78	0.59	0.42	0.62	0.70		

Componentes Principales: Ejemplo



Componentes Principales: Ejemplo



1.4.2.- Selección de Variables

Selección de Variables o Atributos

- Eliminar variables relevantes o dejar algunas irrelevantes, puede alterar los resultados del proceso del análisis.
- El tener más variables de las necesarias, puede hacer más lento el análisis.
- Idea general: reducir el tamaño de la base de datos removiendo las variables irrelevantes o redundantes.
- Esto facilita la interpretación de los resultados obtenidos.

Selección de Variables o Atributos

- Para p atributos, hay 2^p posibles subconjuntos.
- Una búsqueda exhaustiva de las variables es inviable.
- Existen métodos heurísticos que exploran un espacio reducido de variables.
- Estos métodos son ambiciosos, ya que mientras buscan en el espacio correspondiente, buscan la mejor opción en ese momento.
- La estrategia es hacer una selección óptima local, esperando que esto lleve a una solución óptima global.
- Los “mejores” y “peores” atributos son, generalmente, seleccionados en base a tests de significancia estadística que asumen que los atributos son independientes.

Selección de Variables o Atributos

Selección hacia adelante `(forward)`:

- Se parte con un conjunto vacío de atributos y se determina el "mejor" atributo desde los originales, el que pasa a formar parte del "grupo seleccionado".
- En cada paso, se selecciona el mejor de los atributos disponibles.

Eliminación hacia atrás `(backward)`:

- Se parte con un conjunto completo de atributos y se en cada paso se elimina el "peor" de ellos.

Selección de Variables o Atributos

Combinación forward y backward:

- En cada paso, se selecciona el mejor de los atributos y remueve el más malo entre los restantes.

Se establece un criterio de parada, que puede diferir para cada mt́odo.

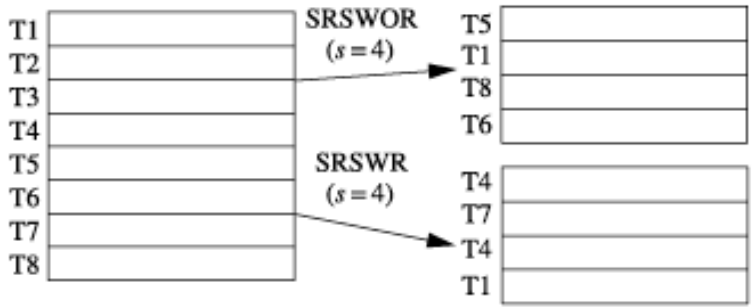
1.4.3.- MUESTREO

Muestreo

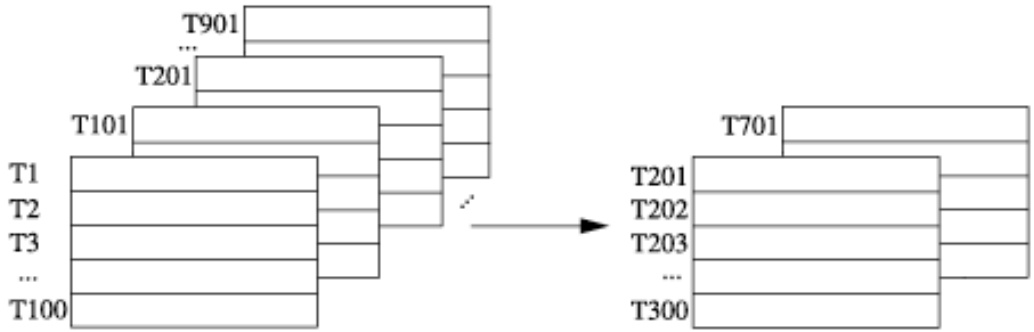
El muestreo se puede utilizar como medida de reducción de datos, ya que permite representar un conjunto de datos grande mediante un sub-conjunto de ésta.

Suponga que tenemos un conjunto de datos, D, que contiene N tuplas o registros.

Muestreo



Cluster sample
$(s = 2)$



Startified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

1.5.- Transformación de Datos

Transformación y Discretización de datos

En esta etapa del preprocesamiento, los datos son transformados o consolidados con el fin de que:

- el proceso de minería sea más eficiente
- los patrones encontrados sean más fáciles de entender

Transformación y Discretización de Datos: Estrategias

1. Suavizamiento
   $\Longrightarrow$ remueve ruido.

2. Construcción de atributos
   $\Longrightarrow$ creación de nuevos atributos.

3. Agregación
   $\Longrightarrow$ se aplican operaciones de resumen o agregación a los datos.

4. Discretización
   $\Longrightarrow$ variables numéricas son reemplazadas por categorías inter-valares o conceptuales.

Transformación y Discretización de datos

5. Generación del concepto jerárquicos para datos nominales
   $\Longrightarrow$ calle - ciudad - región país.

6. Normalización
   $\Longrightarrow$ datos son escalados de manera que pertenezcan a un rango menor de valores.

1.5.1.- Normalización

Normalización

- La unidad de medida utilizada puede afectar los análisis.
- Unidades más pequeñas, llevan a un rango mayor de valores para ese atributo.
- Tiende a entregar a esa variables un mayor efecto o “peso”.
- Independencia de la elección de las unidades de medida.
- Normalización $\Longleftrightarrow$ Estandarización.
- Transformación de los datos para que pertenezcan a [-1,1] o [0,1].
- Especialmente útil en algoritmos de clasificación que involucran redes neuronales o medidas de distancias.

Normalización

Sea A una variable numérica con n valores observados, $a_1, \ldots, a_n$.

Normalización min-max

Realiza una transformación lineal de los datos originales, asignando un valor a_i^* a a_i en un rango $[new\_min_A, new\_max_A]$ de la siguiente manera:

$$a_i^* = \frac{a_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A,$$

preservando las relaciones entre los datos originales.

Normalización: Ejemplo

Suponga que los valores mínimo y máximo para la variable `ingreso` son $12.000 y $98.000, respectivamente, y se está interesado en asignar esta variable a un rango entre 0 y 1. Transforme el valor $73.600 de manera que quede en esta nueva escala.

Normalización

Normalización puntaje-z

Los valores de una variable, A, son normalizados en base a su media ($\bar{a}$) y desviación estándar (S_a), calculando:

$$a_i^* = \frac{a_i - \bar{a}}{S_a}.$$

Es útil cuando no conocemos los valores mínimo y máximo o presencia de valores extremos.

Normalización: Ejemplo

Suponga que la media y desviación estándar para la variable `ingreso` son $54.000 y $16.000, respectivamente.

Normalice el ingreso $73.600 mediante el puntaje-z.

Normalización

Una variación del puntaje-z se obtiene si se reemplaza la desviación estándar por la desviación media de la variable A:

$$S'_A = \frac{1}{n}(|a_1 - \bar{a}| + \ldots + |a_n - \bar{a}|),$$

medida que es más robusta frente a observaciones extremas que la desviación estándar.

Normalización

Normalización por escalamiento decimal

Normaliza los datos moviendo la posición decimal de los valores del atributo A, de la siguiente manera:

$$a_i^* = \frac{a_i}{10^j}.$$

donde j es el menor entero tal que:

$$max|a_i^*| < 1$$

El número de posiciones de movimiento depende del mayor valor absoluto de la variable A.

Normalización: Ejemplo

Suponga que los valores de la variable *A* van desde -986 a 917. Encuentre el valor por el que se debe realizar el escalamiento decimal.

1.6.- Lecturas Sugeridas

Lecturas Sugeridas

- Little, R.J.A., Rubin, D.B. 2002. Statistical analysis with missing data. 2nd Ed. Wiley.
- Mitchell, T. 1997. Machine Learning. McGraw-Hill.