Capítulo 5

Preprocesamiento

Ricardo García Ródenas

Este es un conjunto de tareas encaminadas a la preparación de los datos. La primera tarea es una **exploración visual y estadística de los datos** con el objetivo de detectar y analizar incoherencias. Un aspecto clave es el tratamiento de los llamados **missing values.** Hay que determinar qué se hace con los registros incompletos. Otra tarea es la detección y tratamiento de los **outliers.** Estos registros poseen valores anómalos que pueden influir fuertemente en las conclusiones del análisis. Otro aspecto importante es la magnitud con las que se miden las diferentes variables, pudiendo influir en el análisis de los datos por lo que en ocasiones conviene la **estandarización de datos** para convertirlos todos a una misma escala. Finalmente, si el tamaño de los datos es excesivamente grande, se requiere la **reducción de la dimensionalidad** y/o **selección de variables** significativas. El conjunto de variables se transforma linealmente a un número menor de variables pero manteniendo la máxima información posible. Ejemplos de estas técnicas son el **análisis de componentes principales** y el **análisis factorial** pero no se analizarán en esta capítulo.

5.1. Estadística descriptiva

5.1.1. Medidas de centralidad

En una base de datos existen multitud de registros, y por tanto aparecen multitud de valores que puede tomar una variable. Supongamos que tenemos una base de datos con la altura de una determinada población. En esta base de datos existen personas de todo tipo, altas y bajas. Denotemos $x_1, x_2, x_3, \ldots, x_n$ las n alturas registradas en la base de datos. Las medidas de centralidad persiguen resumir este conjunto de valores en uno único. Para las variables que tienen una naturaleza cuantitativa, como el citado ejemplo, las medidas de centralidad más usada son:

Media (muestral), definida por

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \tag{5.1}$$

Mediana. Supongamos que hemos reordenado los datos de forma creciente (o decreciente) y ahora el nuevo vector lo denotamos por $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$. En función de si n es un número par o impar la mediana se calcula:

$$n \text{ es impar} \Rightarrow Me = x_{\left(\frac{n+1}{2}\right)} \tag{5.2}$$

$$n \text{ es par} \Rightarrow Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \tag{5.3}$$

Cuando los datos tiene naturaleza discreta, por ejemplo la variable x codifica numéricamente el nombre de un país (1 Andorra, 2 Angola, etc.) las anteriores medidas no tienen significado (¿Qué país es 24,67?). Para variables cualitativas una medida de centralidad es la **moda**, que es el valor más repetido.

5.1.2. Medidas de dispersión

Supongamos que tenemos dos máquinas expendedoras de refrescos. Hemos hecho un experimento para conocer cual funciona con mayor exactitud. Hemos rellenado mil vasos con cada una de ellas y obtenemos el mismo valor medio (e igual al que deseamos llenar). Ambas máquinas no llenan exactamente los vasos, sino que van compensando excesos con defectos de refresco. Medir esta variación, como los diversos valores se alejan del valor medio, es lo recogen las medidas de dispersión. La medida más natural sería el error medio:

$$\sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{n} \tag{5.4}$$

donde $|z|$ es el valor absoluto del número z. La expresión $|x_i - \bar{x}|$ representa el **error** de la $i-$ésima medición. Esta cantidad mide la diferencia (sin signo) entre el valor medio y el valor i-ésimo que se ha observado. Recordad que el valor absoluto de un número no negativo es el mismo y se le elimina el signo si es negativo. Por ejemplo $|-7| = |7| = 7$. La función valor absoluto tiene dificultades para su manipulación matemática por ese motivo se ha reemplazado el error en la definición de medidas de dispersión por el el error elevado cuadrado, evitando así el uso del valor absoluto. El resultado son las siguientes medidas de dispersión:

$$\textbf{Varianza muestral} \equiv s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{5.5}$$

$$\textbf{Desviación estándar muestral} \equiv s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \tag{5.6}$$

notar que en su definición se ha dividido por $n-1$ datos en lugar del número de datos $n-1$. Esto se realiza para obtener una mejor estimación del valor que se obtendríamos si realizáramos los cálculos con todos los datos de la población en lugar de los disponibles en la base de datos.

Aunque el valor de la desviación estándar muestral s solo es una aproximación al valor (5.4), facilita el análisis interpretarlo como la desviación media. Siguiendo con el ejemplo de las máquina expendedoras. Las máquinas A y B pueden tener la misma media, por ejemplo, $\bar{x}_A = \bar{x}_B = 33$ cl por lo que ambas estarán bien calibradas, pero diferir en su desviación muestral, por ejemplo $s_A = 1$ cl. y $s_B = 0{,}5$ cl, indicando que en el caso de la máquina A unas veces sus llenados son de 34 cl y otras de 32 cl, mientras que la máquina B su comportamiento es más regular, llenando unas veces 33,5 cl y otras 32,5 cl. La desviación estándar (muestral) nos informa de cuanto alejado (en media) estará una observación x de la media (muestral).

Otra medida de dispersión de los valores que se observan en la muestra es el

$$\textbf{Rango} \equiv R = \max_i \{x_i\} - \min_i \{x_i\} \tag{5.7}$$

Ejemplo 2.1 Estadistica Descriptiva.ipynb Este cuaderno Colab está dedicado a la estadística descriptiva de datos continuos.

Supongamos que los valores de una variable (continua) son $1, 2, 3, \ldots, 10$. Calcula la media, la mediana, la desviación estándar muestral, la varianza muestral y el rango.

Supongamos que tenemos la muestra $1, 2, 3, \ldots, 10$. ¿ Cómo se verían afectados la media, la mediana, la desviación estándar muestral, la varianza muestral y el rango si todos los datos de la muestra se multiplican por 2? ¿Y si en lugar de multiplicarlos por dos se le hubiera sumado 2 a cada uno de ellos?

5.2. Distribución normal

Una variable aleatoria X puede tomar potencialmente varios (infinitos en ocasiones) valores. Por ejemplo el error cometido al evaluar cierta magnitud, en un principio este error puede ser positivo o negativo y arbitrariamente grande. En estadística se proponen modelos para reflejar como varían estos posibles valores, o más precisamente qué probabilidad tiene que el valor de la variable se encuentre en un determinado intervalo. Las variables aleatorias pueden ser continuas (pue-

den tomar todos los valores de un intervalo) o discretas (toman un número finito, como los resultados de un dado, o infinitos numerables $0, 1, 2, 3, \ldots$). La variable aleatoria continua más importante en la distribución normal. La caracterización de una variable aleatoria continua viene definida por dos funciones, **función de densidad de probabilidad** (f.d.p) y **función de distribución acumulada** o simplemente **función de distribución** (F.D.). En el caso de la distribución normal estas son:

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.8}$$

$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \phi_{\mu,\sigma^2}(t)\mathrm{dt} \tag{5.9}$$

La primera observación es que si observamos la definición de la f.d.p aparecen dos parámetros el primero es μ **la media** y el segundo σ^2 que es **la varianza** (σ es la desviación típica). El aspecto de la f.d.p de una variable normal se muestra en la figura 5.1. Esta representación de la distribución normal se le conoce como **campana de Gauss**. Observar que la media μ define por donde pasa su eje de simetría y la varianza σ^2 informa si la campana está mas o menos cerrada entorno a la media. Varianzas pequeñas implican campanas concentradas entorno a la media. De esta observación se deduce que una variable aleatoria normal queda completamente determinada si se conoce sus dos parámetros: media y varianza. Por eso se escribe

$$X \sim N(\mu, \sigma) \tag{5.10}$$

para indicar que X tiene una distribución normal con media μ y varianza σ^2 (o equivalentemente desviación típica σ).

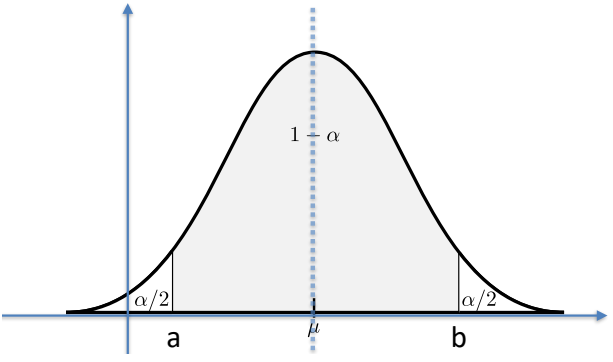

Figura 5.1: Área bajo la curva normal.

5.2.1. Cálculos de probabilidades y tipificación

En una variable aleatoria continua, como lo es la distribución normal, la probabilidad está asociada al área que queda por debajo de la f.d.p. Si quisiéramos calcular la probabilidad que una variable normal estuviera comprendida entre los valores a y b (equivale a calcular el área de la región gris de la figura 5.1) tendríamos que emplear la función de distribución, calculando:

$$P(a \leq X \leq b) = \Phi_{\mu,\sigma^2}(b) - \Phi_{\mu,\sigma^2}(a) \tag{5.11}$$

El problema de la fórmula (5.11) es que la función de distribución Φ solo se puede calcular numéricamente. Históricamente se solventó este problema empleando tablas estadísticas y empleando la tipificación de las variable que permitía transformar una variable aleatoria normal general de parámetros arbitrarios a una variable **normal estándar** (de media igual a 0 y varianza igual a 1) mediante la transformación:

$$Z = \frac{X - \mu}{\sigma} \tag{5.12}$$

Hoy en día, con el uso de los ordenadores, la utilización de las tablas es innecesaria e ineficiente. El cuaderno *Ejemplo 2.2 Distribución Normal.ipynb* recoge un ejemplo de cómo se pueden realizar estos cálculos con Python.

Un cálculo habitual en muchos procedimientos estadísticos es determinar los cuantiles de la distribución normal estándar Z. Un cuantil de orden $0 < p < 1$ es un umbral que deja a una proporción p de valores de la variable por debajo de dicho umbral. Por ejemplo, el primer cuartil (cuantil de orden $p = 0{,}25$) dejaría al $25\,\%$ de los valores de la variable por debajo de dicho valor. Entonces, se tiene que calcular el umbral Z_p de una distribución normal estándar cumpliendo:

$$P(Z \leq Z_p) = p \tag{5.13}$$

Si observamos la figura 5.1 y nos pidieran que calculamos el cuantil de orden $p = 1 - \frac{\alpha}{2}$ diríamos $Z_{1-\frac{\alpha}{2}} = b$ y si fuera el de orden $p = \frac{\alpha}{2}$ contestaríamos $Z_{\frac{\alpha}{2}} = a$

Este cálculo requiere calcular la antiimagen de la función de distribución, esto es:

$$Z_p = \Phi^{-1}(p) \tag{5.14}$$

Notar que cuando trabajamos con la normal estándar omitimos los parámetros en la especificación de la f.d.p o F.D. El cuaderno *Ejemplo 2.2 Distribución Normal.ipynb* recoge un ejemplo de cómo se pueden realizar estos cálculos con Python.

Aunque hoy en día no es necesario la tipificación de las variables normales para efectuar cálculos ya que no es imprescindible recurrir a las tablas de la distribución normal, desde el punto de vista teórico esta operación se sustenta en una importante propiedad que es que la transformación lineal de las variables normales sigue

siendo una distribución normal. Más formalmente si $X \sim N(\mu, \sigma)$ entonces la transformación lineal cumple $aX+b \sim N(a\mu+b, a\sigma)$. Otra propiedad esencial de las variables normales es que $X \sim N(\mu_a, \sigma_a)$ y $Y \sim N(\mu_b, \sigma_b)$ entonces la suma de ellas $X + Y \sim N(\mu_a + \mu_b, \sqrt{\sigma_a^2 + \sigma_b^2})$.

Ejemplo 2.2 Distribución Normal.ipynb Este cuaderno Colab está dedicado a realizar los cálculos involucrados en las variables normales y descritos en esta sección.

Sea $X \sim N(2, 5)$ calcular las siguientes probabilidades:

a) $P(X \geq 0)$

b) $P(-1 \leq X \leq 3)$

c) $P(X \leq 5)$

Sea $Z \sim N(0, 1)$, el cuantil de orden q es el valor Z_q cumpliendo $P(Z \leq Z_q) = q$. Calcula:

a) El primer y tercer cuartil (son los cuantiles de orden $0{,}25$ y $0{,}75$), esto es, $Z_{0,25}$ y $Z_{0,75}$.

b) Los cuantiles de orden $0{,}975$ y $0{,}995$.

Sea $X \sim N(5, 2)$ encuentra el intervalo $[a, b]$ centrado en la media de la variable cumpliendo:

a) $P(a \leq X \leq b) = 0{,}95$

b) $P(a \leq X \leq b) = 0{,}99$

Sea $X \sim N(1, 1)$ e $Y \sim N(-1, 1)$, calcular las siguientes probabilidades:

a) $P(X + Y \leq 1)$

b) $P(X - 2Y \geq 1)$

5.3. Outliers

Los *outliers* son observaciones que tienen valores inusuales, muy grandes o muy pequeños en relación al resto de la muestra. Por ejemplo, supongamos que estamos analizando la distribución de ingresos en una determinada población, y por azares del destino reside una de las principales fortunas. Este registro será *inusualmente* grande. Si la localidad fuera pequeña, este valor podría afectar significativamente a todo el análisis creando la imagen falsa de que todos los habitantes de la localidad fuesen ricos.

Otra fuente para la aparición de *outliers* es la introducción de errores en los datos, por ejemplo, la introducción errónea de un dígito que haría que el número fuese (como mínimo) diez veces más grande de lo que debiese ser.

El primer paso es la detección de estos datos, y tras su análisis, habría que corregirlos, sacarlos del análisis o emplear un método (estadístico) que fuera insensible a este tipo de datos. Veamos varios métodos para detectarlos.

Los métodos se basan en construir bandas (intervalos) en los que deben estar los datos. Si un dato se sale de dicha banda se dictamina que se trata de un *outlier*. Formalamente

$$\text{Si } x_i \notin [x_L, x_U] \Rightarrow x_i \text{ es un outlier} \tag{5.15}$$

Método 1: asumiendo una distribución normal de las variables. Para construir estas bandas suponemos que la variable en cuestión sigue una distribución normal. Además consideramos que las bandas están centradas entorno la media, esto es, $x_L = \mu - k\sigma$ y $x_U = \mu + k\sigma$, por tanto tenemos que determinar un único parámetro k. Calculemos la probabilidad que el dato esté dentro de dicho intervalo

$$\begin{aligned} P\left(x_i \in [\mu - k\sigma, \mu + k\sigma]\right) &= P\left(\frac{x_i - \mu}{\sigma} \in [-k, k]\right) \\ &= P\left(Z_i \in [-k, k]\right) = 1 - \alpha \end{aligned} \tag{5.16}$$

entonces k es el cuantil de una distribución normal tipificada de orden $1 - \alpha/2$. Para calcular dicho valor basta con considerar la distribución acumulada de la distribución normal, denotada por Φ, y calcular la antiimagen de $1 - \frac{\alpha}{2}$, esto es:

$$k = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{5.17}$$

En la construcción de las bandas hay que destacar dos hechos importantes,

- Los valores de los parámetros μ y σ son desconocidos, por lo que se reemplazan por sus estimaciones de la media muestral $\bar{x}$ y desviación típica s.

- Hemos calculado las banda para que la probabilidad de que un dato esté dentro de ella sea $p = 1-\alpha$. Sin embargo, la probabilidad global que n datos independientes lo estén es $p_g = p^n$. Esto tiene la consecuencia que elegir bandas amplias para que la probabilidad de observar un único dato fuera de esas bandas sea casi cero no garantiza que al repetir el proceso con numerosas observaciones alguna se escape. O dicho con un ejemplo, la probabilidad que una persona supere los dos metros de altura es casi cero, pero la probabilidad que en una ciudad haya una persona que supere los dos metros es grande.

  - Vamos a corregir el anterior fenómeno. La probabilidad global (de toda la muestra) que esté fuera de las bandas es $\alpha_g = 1 - p_g = 1 - p^n$. Despejando, obtenemos la siguiente relación:

$$p = \sqrt[n]{1-\alpha_g} \Rightarrow \alpha = 1 - p = 1 - \sqrt[n]{1-\alpha_g} \tag{5.18}$$

    Esta expresión relaciona la probabilidad de que algún punto de la muestra esté fuera de las bandas (α_g) con la probabilidad de un único dato (α). Esto nos permite elegir una probabilidad global para toda la muestra α_g y posteriormente calcular cual debería ser la probabilidad que habría que exigirle a un solo dato de la muestra.

  - Este problema se puede abordar alternativamente empleando el llamado criterio de Chauvenet consistente en tomar un valor α dependiente del número de datos. Este método considera:

$$\alpha = \frac{1}{2n} \tag{5.19}$$

Método 2: Rango intercuartiles. El siguiente método calcula las bandas mediante la expresión:

$$x_L = Q_1 - k(Q_3 - Q_1) \tag{5.20}$$
$$x_U = Q_3 + k(Q_3 - Q_1) \tag{5.21}$$

donde Q_1 y Q_3 son el primer y tercer cuartil de la muestra. Los valores usuales para k son $k = 1{,}5$ (outlier leve) y $k = 3$ (outlier extremo).

Ejemplo 2.3 Outliers.ipynb Este cuaderno Colab está dedicado a implementar los procedimientos para determinar otuliers.

Supongamos que tenemos la muestra $1, 2, 3, 4, \ldots, 10$. Calcula las bandas $[x_L, x_U]$ para determinar la existencia de oultiers empleando los siguientes procedimientos:

a) Aplicando el método 1 y tomando $\alpha = 5\,\%$.

b) Aplicando el método 1, e introduciendo la corrección $\alpha_g = 5\,\%$.

c) Aplicando el rango intercuartiles con $k = 1{,}5$.

d) Aplicando el rango intercuartiles con $k = 3$.

5.4. Método de jackknife: aplicación a determinación de observaciones influyentes

Los procedimientos estadístico, como regresión, clasificación o clustering, pueden estar fuertemente influido por unas determinadas observaciones-registros (las llamadas *observaciones influyentes*) que poseen valores extremos en determinadas variables.

Por ejemplo, si estamos ajustando una recta a una nube de puntos, puede existir un punto muy alejado de la nube que fuerza a que la recta de ajuste intente acercarse a él, alejándose del grupo.

Una técnica estadística para detectar este fenómeno es la *técnica de jackknife*. Supongamos que el método estadístico genera un índice de la calidad del modelo, que denotamos por ψ. Por ejemplo la suma de los errores al cuadrado en regresión o en análisis cluster. Entonces para cada dato i de la muestra se realiza el análisis estadístico sin él, esto permitirá tener una mustra de índices $\{\psi_i\}$. A continuación calculamos los outliers de la muestra $\{\psi_i\}$. Cada *outlier* estará asociado a un registro que produce variaciones insuales del ajuste-calidad del modelo.

Por ejemplo, esta técnica ha sido aplicada por [CM96] en el análisis cluster. Otros métodos estadísticos y gráficos han sido desarrollados por otros autores como [Ste06] para el análisis cluster.

Ejemplo 2.4 Observaciones influyentes.ipynb Este cuaderno Colab está dedicado a implementar los procedimientos para determinar observaciones influyentes en el cálculo de la media y en la mediana.

5.5. Escalamiento (estandarización)

El orden de magnitud de las variables influyen en los procedimientos estadísticos. Por ejemplo, si tenemos una muestra donde se recogen el peso y la altura de ciertos individuos es importante las unidades en cómo se miden las variables. Si la altura la cuantificamos en metros, los individuos posiblemente estarían en el

intervalo $[0, 2]$ mientras que el peso, si fuese medido en kilogramos, en el intervalo $[0, 200]$. Una variable tiene un rango 100 veces más grande que la otra. Este hecho puede afectar al procedimiento. Por ejemplo, si calculásemos una distancia entre individuos, la componente relativa a la altura sería irrelevante frente al peso. Los procedimientos más habituales para el escalamiento de las variables son:

- **Estandarización por rangos**. Se reemplazan las variables por

$$x_i^{new} = \frac{x_i^{old} - \text{mín}(x_i^{old})}{\text{máx}(x_i^{old}) - \text{mín}(x_i^{old})} \tag{5.22}$$

- $Z-$**score estandarización.** Consiste en hacer que las variables tengan media 0 y desviación estándar 1, usando la fórmula de la tipificación

$$x_i^{new} = \frac{x_i^{old} - \overline{x}_i}{s_i} \tag{5.23}$$

  donde $\overline{x}_i$ es la media muestral de la variable x_i y s_i su desviación estándar muestral.

- **Escalamiento decimal**. Se trata de dividir por potencias de 10, esto es:

$$x_i^{new} = \frac{x_i^{old}}{10^m} \tag{5.24}$$

  siendo m el número de dígitos del mayor dato en valor absoluto.

Esta operación se realiza para cada una de las variables, trabajando posteriormente sobre las variables estandarizadas. Notar que si un nuevo dato se incluye posteriormente en el análisis, esté deberá sufrir el mismo procedimiento de estandarización.

[MC88] investigaron ocho métodos diferentes de estandarización para el problema de análisis cluster bajo diferentes condiciones de error, concluyendo que la estandarización por el rango es el método más efectivo, más incluso que el usual *z-score*.

Ejemplo 2.5 Escalamiento de datos.ipynb Este cuaderno Colab está dedicado a los procedimientos para el escalamiento de variables continuas.

Considerar que la variable X toma los valores $1, 2, 3, \ldots, 10$. Se pide:

a Estandarizar la variable mediante rangos. ¿Cuanto vale la media y la desviación estándar muestral de la variable escalada?

b) Repetir el apartado anterior con el escalamiento $Z - score$?

Tabla 5.1: Medidas de correlación entre las variables

	y **continua**	y **categórica**
x_i **continua**	Correlación de Pearson	Fisher score
x_i **categórica**	Fisher score/ ANOVA	Valor información, V de Cramer, ganancia de entropía

5.6. Selección de las variables

Las bases de datos pueden contener miles de variables, causando problemas de computación e interpretación. Un primer problema que hay que abordar es determinar que conjunto de estas variables serán empleadas en el análisis.

En el análisis cluster se suele recurrir a la **reducción de la dimensionalidad** de la matriz de datos X mediante **análisis de componentes principales** que transforma la anterior matriz a una matriz $\widetilde{X}$ de dimensiones mucho menor que la original. Esto es, se mantienen el número de registros (filas) pero se reduce el número de columnas. Estas nuevas variables son transformaciones lineales de las variables originales. Esta técnica la describiremos más adelante.

En problemas de clasificación y regresión se utilizan las medidas de correlación entre la variable respuesta y y el conjunto de regresores (características) x_i quedándose con las que tienen mayores correlaciones. Se intenta considerar sólo aquellas que tienen mayor información de la variable de interés y. Las medidas de correlación intentan medir el grado de dependencia entre dos variables. La medida de este tipo más conocida y aplicada es el **coeficiente de correlación de Pearson** . Este coeficiente se aplica a dos variables cuantitativas y mide la dependencia lineal. Este coeficiente varía entre -1 y 1. Valores próximos a cero indican que ambas variables son independientes, mientras que valores próximos a 1 indican que si la variable x_i aumenta la y lo hará también y de una forma lineal (la relación funcional que liga y y x_i es aproximadamente una recta). Si el valor del coeficiente fuese próximo a -1, aumentos de la variable x_i irían aparejados de descensos (lineales) de la variable respuesta y.

Las variables pueden tener diversa naturaleza, por ejemplo, en problemas de clasificación la variable y tiene una naturaleza cualitativa y ya no es aplicable entonces el coeficiente de correlación de Pearson. Se han desarrolladas medidas alternativas para estas situaciones, intentando emular el comportamiento del coeficiente de correlación de Pearson. La tabla 5.1 resume las diferentes casuísticas e indica la medida de correlación a emplear.

Un analista de datos se enfrenta al siguiente dilema: Emplear todos los datos para aprovechar así toda la información y obtener así modelos más precisos o emplear exclusivamente las variables relevantes. En estadística se ha acuñado el llamado **principio de parsimonia** que apuesta por emplear los modelos más sencillos posibles y no complicar el modelo/análisis si no es significativamente mejor. Notar que si se emplea todos los datos se paga un precio. El más evidente es el incremento

del coste computacional. El segundo inconveniente es su interpretación, no se llega a saber cuales son las variables esenciales del problema. El tercer hecho es que si se va a utilizar el modelo de regresión/clasificación con nuevos registros habrá que recopilar todas las variables del modelo. Por ejemplo si queremos predecir si un cliente de un banco devolverá o no un préstamo personal y como característica introdujimos el color de su coche (un absurdo, pero para ilustrar el hecho) tendremos que conocer el color del coche del nuevo cliente que queremos evaluar para poder usar el modelo.

5.7. Ponderación de variables

El problema anterior aborda cómo elegir un subconjunto significativo de variables. Una vez resuelto, todavía subyace el problema de que no todas la variables son igualmente importantes. Cuando hemos estandarizados las variables las hemos homogeneizado para poderlas comparar.

En análisis cluster una forma de resolver este problema es asignar pesos a las variables en el cálculo de distancia. Algunos métodos de ponderación para el análisis cluster son discutidos en [AGK82]. En problemas de regresión/clasificación son los propios métodos quien llegan a ponderar implícitamente las variables estandarizadas.

Bibliografía

[AGK82] D. Art, R. Gnanadesikan, and J. Kettenring. Data-based metrics for cluster analysis. *Utilias Mathematica*, (21A):75–99, 1982.

[CM96] R. Cheng and G.W. Milligan. Measuring the influence of individual data points in cluster analysis. *Journal of Classification*, (13):315–335, 1996.

[MC88] G.W. Milligan and M.C. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, (5):181–204, 1988.

[Ste06] D. Steinley. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34, 2006.