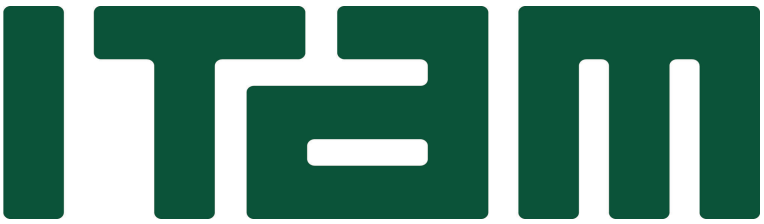
Instituto Tecnológico Autónomo de México



Clasificación no supervisada utilizando modelos bayesianos

TESIS

QUE PARA OBTENER EL TÍTULO DE

Lic. Actuaría

PRESENTA

Montserrat Vizcayno García

ASESOR

Dr. Juan Carlos Martínez Ovando

MÉXICO, D.F. 2017

"Con fundamento en los artículos 21 y 27 de la Ley Federal de Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**Clasificación no supervisada utilizando modelos bayesianos**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la biblioteca Raúl Baillères Jr., autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una prestación"

Montserrat Vizcayno García

Fecha

Firma

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Estructura de la Tesis

Capítulo 2

Inferencia Bayesiana

En inferencia estadísitca, hay dos enfoques que prevalecen en la practica al momento de interpretar la probabilidad: La inferencia Bayesiana y la inferencia frecuentista. Estos dos paradigmas inferenciales, suelen diferir con respecto a la naturaleza fundamental de la probabilidad. La alternativa frecuentista la define de una manera más restrictiva, como el límite de la frecuencia relativa de un evento en un gran número de intentos, en el contexto de que dichos experimentos sean aleatorios y bien definidos. Por otro lado, la inferencia Bayesiana, es capaz de asignar probabilidades a cualquier evento, aún cuando no hay un proceso aleatorio de pormedio, se puede decir que en este caso, la probabilidad es una manera de representar el nivel de creencia

sobre un evento, o dada una evidencia

Por lo que el proceso de aprendizaje comprendido en la Inferencia Bayesiana, consiste en modificar las creencias iniciales de los parámetros antes de observados los datos, por un conocimiento posterior actualizado, que combine tanto el conocimiento previo como la información disponible.(Hall, 2012).

En las siguientes secciones se hablará sobre los conceptos básicos para poder llevar a cabo el proceso de inferencia estadística bajo el enfoque bayesiano.

2.1. Verosimilitud

En general, para el análisis estadístico de datos observados, $x_1, ..., x_n$, con $x_j \in \mathbb{R}^p$, se supone un modelo estocástico de n-variables aleatorias independientes e identicamente distribuidas, $(X_j, j = 1, ..., n)$.

Para efectos prácticos, se define un modelo paramétrico $P(X) = F(x|\theta)$, donde $F(\cdot|\theta)$ es una función de distribución dada [1]. En esta ocasión se trabaja con el segundo caso, en el cual θ es el parámetro que indiza dicha distribución y toma valores en el espacio parametral $\Theta \in \mathbb{R}^p$ (con $p < \infty$); es decir, que

[1] ya sea por una función de masa de probabiblidad en el caso discreto, o por una función de densidad en el caso continuo

hay un número finito de parámetros [2]. De manera que, bajo el supuesto de independencia, tenemos lo siguiente:

$$P(X_1, ..., X_n) = \prod_{j=1}^{n} P(X_j) \tag{2.1}$$

donde $X_j = (X_{j1}, ..., X_{jp})$, es un vector p-dimensional.

Mediante el método de máxima verosimilitud se utiliza la información disponible del conjunto de datos, para encontrar los valores más probables del parámetro θ dentro del espacio parametral Θ, asociados con los datos $X_1, ..., X_n$ observados.

Al considerar $P(X_j)$ como una función de distribución o densidad, de las observaciones dado el parámetro θ, $P(X_j) = F(\cdot|\theta)$, obtenemos la siguiente función de verosimilitud:

$$L(x_1, ..., x_n; \theta) = f(x_1, ..., x_n|\theta) = \prod_{j=1}^{n} f(x_j|\theta) \tag{2.2}$$

de donde se sigue para obtener el valor de θ que maximiza la función,

$$\hat{\theta} = argmax \prod_{j=1}^{n} f(x_j|\theta) \tag{2.3}$$

y así encontrar el Estimador de Máxima Verosimilitud (EVM)

[2] al contrario del caso no paramétrico, donde el número de parámetros es infinito ($\Theta \in \mathbb{R}^{\infty}$)

2.2. Paradigma Bayesiano

El Paradigma Bayesiano se basa en el aprendizaje, por lo que, los datos añaden nueva información al conocimiento previo y de esta forma, se actualizan las creencias sobre los parámetros de interés. Por lo que se deben especificar las creencias anteriores al análisis de los datos y asignar una distribución inicial $\pi(\theta)$ definida como una medida de probabilidad sobre Θ, que describe el comportamiento de θ con la información disponible sobre este parámetro antes de haber observado los datos. Esta información es entonces combinada con los datos para producir la distribución a posteriori o final, $\pi(\theta|X_1, .., X_n)$, que expresa lo que se conoce de los parámetros, una vez que se introdujeron los datos.

Se utiliza el teorema de Bayes como un mecanismo para combinar la información a priori, $\pi(\theta)$, con la información proporcionada por los datos, $P(X_1, ..., X_n|\theta)$, que como se mencionó anteriormente, esta última es la función de verosimilitud.

$$\pi(\theta|X_1, ..., X_n) = \frac{P(X_1, ..., X_n|\theta) \cdot \pi(\theta)}{P(X_1, ..., X_n)} \tag{2.4}$$

donde el denominador, $P(X_1, ..., X_n) = \int_\Theta P(X_1, ..., X_n|\theta) \cdot \pi(\theta) d\theta$, es una integral sobre todos los valores de θ del producto de la función de verosimilitud

y la previa del parámetro θ y se toma como una constante de normalización para asegurar que $\pi(\theta|X_1, ..., X_n)$ sea una función densidad propia.

Simplificando, el teorema de Bayes puede ser expresado de la siguiente manera,

$$\pi(\theta|X_1, ..., X_n) \propto P(X_1, ..., X_n|\theta) \cdot \pi(\theta) \tag{2.5}$$

donde $\propto$ denota una relación de proporcionalidad, dada por $P(X_1, ..., X_n)^{-1}$

2.3. Intercambiabilidad

Al relajar el supuesto de independencia, se introduce el concepto de **intercambiabilidad**, el cual reconoce que el orden de las observaciones es invariante ante permutaciones de sus índices; es decir, toda la informacion relevante está contenida en los valores de las X_i's, de forma que sus índices no proporcionan informacion alguna. Observese que el concepto de intercambialilidad generaliza el de independencia condicional: un conjunto de observaciones independientes identicamente distribuidas son siempre un conjunto de observaciones intercambiables (Bernardo, 1998).

Entonces, $\{X_j\}_{j=1}^{\infty}$ son intercambiables si para todo $n < \infty$,

$$P(X_1 = x_1, ..., X_n = x_n) = P(X_{\sigma(1)} = x_1, ..., X_{\sigma(n)} = x_n) \tag{2.6}$$

donde $\{\sigma(1), ..., \sigma(n)\}$ es cualquier permutación de {1,2,...,n}.

2.4. Teorema de Representación

Utilizando en concepto de intercambiabilidad, de Finetti demuestra su famoso teorema de representación para variables dicotómicas. En este caso en particular, la intercambibilidad identifica las observaciones como una muestra aleatoria de un modelo probabilístico específico (Bernoulli) y garantiza la existencia de una distribución inicial sobre su parámetro. En el caso general, para variables aleatorias de cualquier rango y dimensión, la intercambiabilidad identifica las observaciones como una muestra aleatoria de algún modelo probabilístico y garantiza la existencia de una distribución inicial sobre el parámetro que lo describe (Bernardo, 1998).

Teorema de representación de Finetti. Si $\{X_j\}_{j=1}^{\infty}$ son variables aleatorias intercambiables, entonces existe un objeto estocástico θ, tal que:

$$P(X_1, ..., X_n) = \int_{\Theta} \prod_{j=1}^{n} P(X_j|\theta)\pi(\theta)d\theta, \tag{2.7}$$

donde $\theta \in \Theta$ se define como el límite (cuando $n \longrightarrow \infty$) de una función de las X_j's y $\pi(\theta)$ es la función de distribución inicial sobre Θ.

En otras palabras, si la secuencia de observaciones es intercambiable, cualquier subconjunto de éstas, es una muestra aleatoria de un modelo $P(X_j|\theta)$ y existe una distribucion inicial $\pi(\theta)$ que describe la información inicial disponible del parámetro θ.

2.5. Distribución predictiva

Ahora, al utilizar argumentos probabilísticos, como es el supuesto de independencia, se busca obtener la distribución predictiva. Esta distribución es la que define como será el comportamiento de una nueva observación X_{n+1} con base en los datos observados $X_1, ..., X_n$.

$$
\begin{aligned}
P(X_{n+1}|X_1 = x_1, ..., X_n = x_n) &= \frac{P(X_1, ..., X_n, X_{n+1})}{P(X_1, ..., X_n)} \\
&= \frac{\prod_{j=1}^{n} P(X_j) \cdot P(X_{n+1})}{\prod_{j=1}^{n} P(X_j)} \qquad (2.8) \\
&= P(X_{n+1}) \\
&= F(X_{n+1}|\theta) \quad ?
\end{aligned}
$$

Como se observa en (2.8), este supuesto tiene como consecuencia un problema de predictibilidad, ya que la distribución de la nueva observación no estaría tomando en cuenta la información proporcionada por las observaciones anteriores. Por lo que, tomando el estimador de máxima verosimilitud (EMV) se recurre a lo siguiente:

$$
P(X_{n+1}|X_1, ..., X_n) \approx P[X_{n+1}|\hat{\theta}(X_1, ..., X_n)] \qquad (2.9)
$$

Sin embargo, la ecuación (2.9) no puede ser considerada como una distribución predictiva, debido a que estimador del parametro θ es un valor que se obtuvo con respecto a la información de $(X_1, ..., X_n)$ y X_{n+1} es independiente de estas observaciones anteriores.

Por lo que es momento de retomar los conceptos de la sección anterior, y proceder a obtener la distribución predictiva bajo el enfoque bayesiano.

$$\begin{aligned} P(X_{n+1}|X_1,...,X_n) &= \frac{P(X_{n+1}|X_1,...,X_n,X_{n+1})}{P(X_1,...,X_n)} \\ &= \frac{\int_\Theta \prod_{j=1}^n P(X_j|\theta)\cdot P(X_{n+1}|\theta)\cdot\pi(\theta)d\theta}{\int_\Theta \prod_{j=1}^n P(X_j|\tilde{\theta})\cdot\pi(\tilde{\theta})d\tilde{\theta}} \\ &= \int_\Theta P(X_{n+1}|\theta)\cdot\frac{\prod_{j=1}^n P(X_j|\theta)\cdot\pi(\theta)d\theta}{\int_\Theta \prod_{j=1}^n P(X_j|\tilde{\theta})\cdot\pi(\tilde{\theta})d\tilde{\theta}} \\ &= \int_\Theta P(X_{n+1}|\theta)\cdot\pi(\theta|X_1,...,X_n)d\theta \end{aligned} \tag{2.10}$$

donde $\pi(\theta|X_1,...,X_n)$ es la distribución final o posterior de θ dado $X_1,...,X_n$, que se define como la creencia sobre el comportamiento de θ después de observar los datos. A su vez, la ecuación (2.10), también se puede ver de la siguiente manera,

$$P(X_{n+1}|X_1,...,X_n) = E_{\theta|X_1,..,X_n}[P(X_{n+1}|\theta)] \tag{2.11}$$

Como la esperanza de la distribución $X_{n+1}|\theta$, condicional a los datos observados, $(X_1 = x_1,...,X_n = x_n)$. De esa manera la predicción del comportamiento

de la nueva información considera la información proporcionada por los datos observados anteriormente.

Para llegar a este punto de definir una distribución predictiva, es necesario como primer paso, encontrar la distribución inicial como se describe a continuación.

2.6. Distribución inicial o previa

La distribución inicial o previa, $\pi(\theta)$, se define como la medida de probabilidad sobre Θ que describe el comportamiento del parámetro θ, con base en el nivel de información disponible antes de observardos los datos.

Existen distintas maneras de obtener la distribución inicial para θ, en algunas situaciones es posible basarse en información que proviene de evidencia acumulada de experimentos pasados. De igual manera, la previa puede determinarse de manera subjetiva con base en la experiencia de un experto. En el caso de no contar con información disponible, se puede recurrir a una distribución previa no informativa, la cual expresa información de tipo objetivo sobre el parámetro, como por ejemplo, θ toma valores positivos (Congdon, 2007). En la práctica, hay métodos que resultan más convenientes matemáticamente

hablando, como es el caso de escoger una distribución inicial conjugada, debido a que de esa forma la districión final $\pi(\theta|\cdot)$ es de una forma paramétrica conocida.

Se dice que, $\pi(\theta)$ y $\pi(\theta|\cdot)$ son distribuciones conjugadas, cuando la distribución final pertence a la misma familia que la distribución inicial. Por ejemplo, si la función de verosimilitud es de la familia Gaussiana, entonces al elegir una distribución Gaussiana como distribución previa del parámetro, la media en este caso, nos asegurará que la distribución final se una distribución Gaussiana. En general todas las distribuciones de probabilidad de la familia exponencial cuentan con distribuciones previas conjugadas (Hall, 2012).

2.7. Distribución final o posterior

La distribución posterior se obtiene aplicando el teorema de Bayes. Se combina la información inicial del parámetro $\theta = (\theta_1, ..., \theta_q)$, mediante la distribución $\pi(\theta) = \pi(\theta_1, ..., \theta_q)$ donde $(\theta \in \Theta \subseteq \mathbb{R}^q)$, y la distribución de las observaciones, mejor conocida como función de verosimilitud,

$$P(X_1, ..., X_n|\theta) = \prod_{j=1}^{n} P(X_j|\theta)$$

en la que se dice que las X_j's son intercambiables y condicionalmente independientes dado θ. Por lo que la distribución final de θ es:

$$\pi(\theta_1, ..., \theta_q | X_1, ..., X_n) = \frac{\prod_{j=1}^{n} P(X_j|\theta_1, ..., \theta_q) \cdot \pi(\theta_1, ..., \theta_q)}{\int_\Theta \prod_{j=1}^{n} P(X_j|\tilde{\theta_1}, ..., \tilde{\theta_q}) \cdot \pi(\tilde{\theta_1}, ..., \tilde{\theta_q}) d\tilde{\theta_1}, ..., \tilde{\theta_q}} \tag{2.12}$$

donde el denominador se define como la constante de normalización C tal que $\pi(\theta_1, ..., \theta_q | X_1, ..., X_n)$ sea una densidad propia; es decir,

$\int_\Theta \frac{1}{C} \cdot \prod_{j=1}^{n} P(X_j|\theta) \cdot \pi(\theta) d\theta = 1$. En la práctica, se conoce explícitamente sólo en términos del numerador (Congdon, 2007),

$$\pi(\theta_1, ..., \theta_q | X_1, ..., X_n) \propto \prod_{j=1}^{n} P(X_j|\theta) \cdot \pi(\theta) d\theta \tag{2.13}$$

El paso siguiente sería el cálculo de la distribución final, que puede derivar en dos escenarios:

a) Distribuciones conjugadas; es decir, la distribución previa y la distribución posterior son la misma forma funcional, se facilita el cálculo, ya que la distribución final resulta ser de una forma paramétrica conocida.

b) No conjudadas, donde la distribución final puede no conocerse explícitamente debido a la imposibilidad de obtener la constante de normalización correspondiente.

En este último, es necesario recurrir a distintos métodos de aproximación para calcular la distribución final de θ.

2.7.1. Métodos de Aproximación

El problema de cálculo en la inferencia Bayesiana se reduce a la resolución de integrales que no tienen una solución analítica viable, por lo que se requieren algoritmos de aproximación númerica. Existen distintos tipos de métodos de aproximación como son la aproximación de Laplace o el método Monte Carlo vía Cadenas de Markov (MCMC) (Hall, 2012).

I. Monte Carlo vía Cadenas de Markov

Para la aproximación de integrales, se utilizan en técnicas de simulación estocástica, empleando T simulaciones de los datos para aproximar dicha integral que en su caso es una función de distribución.

Si las simulaciones $(\theta_1^{(t)}, ..., \theta_q^{(t)})$, con $(t = 1...T)$ son independientes e identicamente distribuidas (iid), entonces partiendo de la ecuación 2.11 tenemos que:

$$E_{\theta|X_1,..,X_n}(P(X_{n+1}|\theta)) \approx \frac{1}{T}\sum_{t=1}^{T} P(X_{n+1}|\theta_1^{(t)}, ..., \theta_q^{(t)}) \qquad (2.14)$$

donde la aproximación de 2.14, se puede interpretar como el promedio empírico de $P(X_{n+1}|\theta)$.

A este método se le conoce como Monte Carlo, sin embargo, en el contexto bayesiano, distribuciones como la final o posterior, $\pi(\theta_1, ..., \theta_q|X_1, ..., X_n)$ llegan a ser muy complejas.

Debido a esta problematica, es necesario relajar el supuesto de *iid* de Monte Carlo de manera que $\theta_1^{(t)}, ..., \theta_q^{(t)}$ dependa de $\theta_1^{(t-1)}, ..., \theta_q^{(t-1)}$; es decir, que el valor de $\theta_1^{(t)}, ..., \theta_q^{(t)}$ va a estar influenciado por la información que proporcione $\theta_1^{(t-1)}, ..., \theta_q^{(t-1)}$ y así sucesivamente para cada una de las T simulaciones. Esto se logra mediante una cadena de Markov, con una distribución de transición:

$$g(\theta^{(t)}|\theta^{(t-1)}, X_1, ..., X_n) \tag{2.15}$$

La unión del método de cadenas de Markov con el método de Monte Carlo, es a lo que se le denomina método Monte Carlo vía Cadenas de Markov (MCMC).

La cadena de Markov es un proceso estocástico $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...\}$, donde $\theta^{(i)}$ es el estado del proceso en el tiempo i, y se define como una variable aleatoria cuyos valores se encuentran en un espacio de estados $\Theta \subset I\!R^p$.

Estas cadenas de Markov cumplen con una propiedad en la cual el próximo

estado (t+1) depende solamente del estado actual (t) y no de los estados anteriores.

La idea central de este método es construir una cadena de transición definida por el *kernel* de transición que tenga a la distribución objetivo $\pi(.)$, como la distribución invariante; es decir, que si $\theta^{(t)} \sim \pi$ implica que $\theta^{(t+1)} \sim \pi$. Donde el *kernel* $K : \Theta \times B(\Theta) \rightarrow [0, 1]$, es la función de transición de los estados, que denota la probabilidad de transición,

$$K(\theta, \theta') = P[\theta^{(t+1)} = \theta' | \theta^{(t)} = \theta]$$

$\forall(\theta, \theta') \in \Theta$ entre las iteraciones t y t+1.

Se dice que $\pi(.)$ es la distribución invariante de la cadena de Markov, si se define un *kernel* de transición que satisfaga la condición de balance $K(\theta, \theta')\pi(\theta) = K(\theta', \theta)\pi(\theta')$ (Martínez Ovando, 2004).

De igual manera, esta cadena deberá cumplir con ciertas condiciones de regularidad (Congdon, 2007):

a) Irreductible, si para cualquier pareja de estados $(\theta^{(t)}, \theta^{(s)}) \in \theta$ existe una probabilidad distinta de que la cadena se puede mover de $\theta^{(t)}$ a$\theta^{(s)}$ en un número finito de pasos.

b) Aperiódica. Un estado tiene una periodicidad k, $(k > 1)$, si puede ser

revisitado solamente después de un número de pasos que sea múltiplo de k, de otra manera el estado es aperiódico. por lo que si todos los esados son aperiodicos, la cadena es aperiódica.

c) Recurrencia positiva. Si el número de pasos para visitar cualquier estado de una cadena tiene media positiva.

d) Ergódica. Se asegura que el promedio aritmético (2.14) converja, casi seguramente, con la esperanza calculada bajo la distribución invariante conforme T se va haciendo mas grande ($T \longrightarrow \infty$). Si la cadena cumple con las condiciones anteriores, se dice que tiene ergodicidad.

En la siguiente sección se describe un método para construir cadenas de Markov que cumple con estas características.

II. Gibbs Sampler

Un algoritmo en particular de cadenas de Markov que ha sido útil para problemas multidimensionales es Gibbs sampler, que esta definido en terminos de subvectores de θ.

Supongamos que el parametro vector θ esta dividido en subvectores, $\theta = (\theta_1, ..., \theta_q)$. Cada una de las T iteraciones del Gibbs sampler, compuestas por

q pasos, recorren los subvectores de θ haciendo que cada subconjunto sea condicional al valor de todos los demás. En cada una de estas iteraciones t, se elige un ordenamiento de los q subvectores de θ y a su vez, se genera una muestra de cada θ_j^t a partir de la distribución condicional dado los demás componentes de θ (Gelman et al., 2014).

Por lo que, método Gibbs Sampler, implica hacer una actualización parámetro a parámetro de cada uno de los componentes $\theta_1^{(t)}, ..., \theta_q^{(t)}$, mediante un muestreo sucesivo de las distribuciones condicionales, que al completarse nos da la transición de $\theta^{(t-1)}$ a $\theta^{(t)}$.

$$\textbf{1).}\ \theta_1^{(t)}|\theta_2^{(t-1)}, ..., \theta_q^{(t-1)}, X_1, ..., X_n \sim \prod_{j=1}^{n} P(X_j|\theta_1, \theta_2^{(t-1)}, ..., \theta_q^{(t-1)}) \cdot \pi(\theta_1|\theta_2^{(t-1)}, ..., \theta_q^{(t-1)});$$

$$\textbf{2).}\ \theta_2^{(t)}|\theta_1^{t}, \theta_3^{(t-1)}, ..., \theta_q^{(t-1)}, X_1, ..., X_n \sim \prod_{j=1}^{n} P(X_j|\theta_1^{k}, \theta_2, \theta_3^{(t-1)}, ..., \theta_q^{(t-1)}) \cdot \pi(\theta_2|\theta_1^{t}, \theta_3^{(t-1)}, ..., \theta_q^{(t-1)});$$

.

.

$$\textbf{q).}\ \theta_q^{(t)}|\theta_1^{t}, ..., \theta_{q-1}^{(t)}, X_1, ..., X_n \sim \prod_{j=1}^{n} P(X_j|\theta_1^{(t)}, ..., \theta_{q-1}^{(t)}, \theta_q) \cdot \pi(\theta_q|\theta_1^{(t)}, ..., \theta_{q-1}^{(t)}) \tag{2.16}$$

Este algoritmo genera una secuencia de números autocorrealacionados que

cumplen con las condiciones de regularidad, que eventualmente olvida los valores iniciales, $\theta_1^{(0)}, ..., \theta_q^{(0)}$, usados para la cadena y que termina por converger en una distribución estacionaria. De manera que, $\{(\theta_1^{(t)}, ..., \theta_q^{(t)})\}_{t=1}^{t}$ se define como una cadena de Markov (Congdon, 2007).

Capítulo 3

Clasificación

3.1. Modelos de Clasificación

Los tipos de clasificación se dividen en dos grandes grupos, la clasificación supervisada y la no supervisada. Para efectos de este análisis se utilizará la clasificación no supervisada, debido a que la información no tiene clases previas en las cuales se puedan ubicar las observaciones.

El problema de la clasificación no supervisada consiste en 'adivinar' o encontrar el número de grupos J con $J = 1, ..., n$ en los cuales se pueden clasificar las observaciones. Con base en los grupos definidos, se crean reglas de aso-

ciación con las cuales X_i es asignado a una clase $C_j \in \{1, 2, ..., J\}$

Dichas reglas de asignación pueden definirse bajo dos corrientes:

En la escuela tradicional, la asignación de las observaciones X_i a una clase C_j se define con base en distancias.

$$X_i \in C_{j^*} \text{ si y sólo si } j^* = \text{argmin}_j \, d(X_i, C_j).$$

Por ejemplo, para clasificar un conjunto de datos $X_1, ...X_n$ se encuentran dos grupos (J=2) y se define la regla de asignación como el mínimo de la distancia que hay del punto X_i al centro del grupo uno (C_1) y la distancia que hay del punto X_i al centro del grupo dos (C_2).

$$X_i \in C_1 \text{ si y sólo si } \{d(X_i, C_1) < d(X_i, C_2)\}$$

donde C_1 y C_2 son los centroides de cada grupo.

Este enfoque funciona siempre y cuando los datos sean escalares, pero en caso de que sean de tipo catégorico surge un problema, ya que no hay manera de medir las distancias.

Por otro lado, ha surgido una nueva escuela en la que las reglas de asignación se hacen con base en la probabilidad de que X_i pertenezca a C_j; por lo que los atributos pueden ser escalares, categóricos o texto. $X_i \in C_{j^*}$ si y sólo si

$j^* = \text{argmax}_{j \in J} P(X_i \in C_j)$

Para comenzar la clasificación, se suponen J grupos preexistentes en los cuales al menos una observación X_i se encuentra en cada uno de estos grupos $C_1, .., C_J$. Cabe destacar que cada una de las observaciones X_i's sólo pueden pertenecer a un sólo grupo a la vez).

$$C_j = \{X_i : X_i \sim F(\cdot|\theta_j)\}$$

donde $\theta_j = T_{F_j}(X)$, definiendo a $T_{F_j}(X)$ como un atributo, por mencionar algunos ejemplos, la media, la varianza, la mediana, etc.

Definimos e_j como el número de observaciones dentro del grupo C_j y se tiene que $P(X_i \in C_j) \approx \frac{e_j}{n}$

Bajo este enfoque, surgen varias problemáticas a resolver, que deben ser consideradas al momento de escoger el modelo de clasificación.

Primero, al ser un problema de clasificación no supervisada, se desconoce el número K de grupos o clases en los que se clasificaran los datos, asi que también se desconoce cuáles son X_j's pertencen a la clase C_j. Los parámetros $\theta_1, ...\theta_j$ asociados con $F_1(\cdot|\theta_1), ..., F_j(\cdot|\theta_j)$, son desconocidos, por lo que será necesario estimarlos.

En la próxima sección se propone un modelo, bajo el cual se abordan estas problemáticas.

3.1.1. Modelo basado en Mezclas de distribuciones

La estructura del modelo tipo mezcla aparece debido a que, como es común en los problemas de clasificación no supervisada, no cuenta con información sobre la pertenencia de cada observación a una subpoblación o clase específicas. Por lo tanto, se asume que cada una de las X_i's puede tener una distribución f_j con probabilidad p_j.

Dependiendo del escenario que se plantee, la meta puede ser reconstruir los grupos a los que pertenecen las observaciones para proveer de estimadores para los parámetros de los diferentes grupos, o incluso, estimar el número de grupos.

Las distribuciones mixtas pueden contener un modelo finito o infinito de componentes, que posiblemente pueden ser de distintos tipos de distribuciones, y describen distintas características de los datos.

En este caso, nos enfocaremos al modelo de mezclas con un número finito de

componentes, que se define así,

$$P(X) = \sum_{j=1}^{J} p_j f_j(X|\theta_j) \tag{3.1}$$

donde p_j es la probabilidad de pertenecer al componente o clase C_j, y

$$\sum_{j=1}^{J} p_j = 1$$

Sin embargo, la manera en como esta representado el modelo en 3.1 vuelve complicado derivar el estimador de máxima verosimilitud (cuando existe) y los estimadores bayesianos.

Por ejemplo, si se considera el caso de n observaciones iid $X = (X_1, ...X_n)$ y definimos $p = (p_1, ...p_J)$ y $\theta = (\theta_1, ..., \theta_J)$, vemos que aunque se hayan utilizado previas conjugadas para cada parámetro, para obtener la dostribución previa de manera explícita, requiere que se realice la expanción de la verosimilitud

$$L(\theta, p|x) = \prod_{i=1}^{n} \sum_{j=1}^{J} p_i f_j(x_i|\theta_j) \tag{3.2}$$

en k^n términos, que para la práctica resulta ser muy costoso computacionalmente hablando, por lo que se proponen diferentes enfoques abordar el problema como la introducción de variables latentes que se explica en la siguiente sección (Marin et al., 2005).

I. Variables Latentes

Una manera de facilitar la estimación es introducir, dentro del modelo, variables aleatorias no observadas (variables latentes) $\underline{z} = (z_1, ..., z_n)$, que identifican a que componente j, (j=1,...J) pertenece cada una de las observaciones $x = (x_1, ..., x_n)$.

$$X_i | Z_i = z \sim f(x|\theta_z) \tag{3.3}$$

Donde $Z_i \sim M_j(1; p_1, ...p_J)$. Si tomamos la ecuación 3.1 y definimos a $\pi(\theta, p)$ como la distribución inicial de (θ, p). La distribución posterior es la siguiente,

$$\pi(\theta, p|X) \propto (\prod_{i=1}^{n} \sum_{j=1}^{J} p_i f_j(X_i|\theta_j))\pi(\theta, p) \tag{3.4}$$

Definimos Z como el conjunto de todos los J_n vectores de asignación z, que podemos descomponer en una partición de J conjuntos. Para un vector de asignación dado $(n_1, ..., n_J)$ donde $n_1 + ... + n_J = n$ definimos el conjunto,

$$Z_i = \{\underline{z} : \sum_{i=1}^{n} \mathbb{1}_{z_i=1} = n_1, ..., \sum_{i=1}^{n} \mathbb{1}_{z_i=J} = n_J\}$$

que consiste en todas las asignaciones dadas por una partición, que en este caso es determinada por el vector de asignación $(n_1, ..., n_J)$. Existe un número

r de soluciones enteras no negativas de las n observaciones en las J clases, en las que se cumple que $\sum_{j=1}^{J} n_j = n$.

$$r = \binom{n+k-1}{n} \tag{3.5}$$

De esa manera tenemos la partición $Z = \cup_{i=1}^{r} Z_i$. Y aunque el número total de elementos de Z no sea manejable en términos computacionales J^n, el número de conjuntos de particiones es mucho más manejable al ser del orden $\frac{n^{k-1}}{(k-1)!}$

Ahora, la distribución posterior se puede descomponer de esta manera,

$$\pi(\theta, p|x) = \sum_{i=1}^{r} \sum_{\underline{z} \in Z_i} w(\underline{z}) \pi(\theta, p|x, \underline{z}) \tag{3.6}$$

donde $w(\underline{z})$ se define como la probabilidad posterior, dada la asignación $\underline{z}$. Esta descomposición hace que la distribución posterior le asigne una probabilidad posterior $w(\underline{z})$ a cada posible asignación $\underline{z}$ de los datos, para luego construir la distribución posterior de los parámetros, condicional a esa asignación (Marin et al., 2005).

3.1.2. Label Switching Problem

El término "*label switching*"se utiliza para descrobor la invarianza de la función de verosimilitud al momento de reetiquetar a los componentes del mo-

delo de mezclas. En otras palabras, para cualquier permutación σ de 1,...,k, se define la permutación correspondiente al parámetro θ como,

$$\sigma(\theta) = ((\pi_{\sigma(1)}, ..., \pi_{\sigma(k)}), (\theta_{\sigma(1)}, ..., \theta_{\sigma(k)}))$$

Lo que a continuación nos lleva a la raíz del problema de *label switching*. La función de verosimilitud es la misma para todas las permutaciones de θ. De igual manera, bajo el enfoque bayesiano, si no se cuenta con la información previa que distinga entre los componentes del modelo de mezclas, la distribución previa $\pi(\theta)$ será la misma para todas las permutaciones y por consiguiente la distribución posterior sera simétrica. Dicha simetría puede causar problemas cuando se busca estimar algún atributo relacionado a los componentes del modelo de manera individual. Por ejemplo, gracias a esta simetría, las funciones de densidad predictivas son las mismas para cada componente, de forma que las probabilidades de clasificación no son se utilidad para la clasificación de las observaciones en grupos, ya que son las mismas para cada observación $(1/k)$. De manera similar, al tener la misma distribución posterior, la media de un parámetro dentro de un componente en específico será la misma que la de media de ese parámetro en los demás

componentes del modelo, por lo que en general suele ser una estimación muy pobre para esos parámetros. (Stephens, 2000)

Capítulo 4

Aplicación práctica

4.1. Objetivo

Obtener una clasificación de los clientes de la empresa de empeño en cuestión con base en los modelos escritos en los capítulos anteriores, utilizado la informacióón contenida en las variables de la base de datos, de modo que se les pueda ofrecer distintos productos.

4.2. Descripción de la información

Los datos de la base provienen de una empresa de empeño y microcréditos que comenzó en febrero del 2006, con diez sucursales en el Estado de México y Querétaro.

A diferencia de otras empresas del ramo, ésta ofrece tres esquemas de pago entre los cuales el cliente puede elegir, según le resulte más conveniente:

1. **Tradicional** es la clásica forma de pago en la cual se pagan interés y se cuenta con 5 refrendos, al llegar al último refrendo se tiene que pagar el monto total del préstamo.

2. **Pagos Fijos** consiste en dividir el monto de la deuda más los intereses entre el número de semanas o meses que tiene el cliente como plazo para liquidar la deuda.

3. **Flexible** es una mezcla de los dos esquemas anteriores, consiste en ir pagando intereses o capital según le convenga hasta cubrir el monto total de la deuda en un plazo acordado.

El inventario de objetos que son admitidos para realizar un empeño, se clasifica de la siguiente manera:

Metales

- oro
- plata

Electrónicos

- Televisores
- Minicomponentes
- Celulares
- Dvd´s
- Consolas de video juegos
- Computadoras
- Camaras digitales
- Reproductores de mp3

Otros

- Relojes

4.2.1. Descripción de la base

La base está compuesta por 29822 clientes. Son 11 variables categóricas y 8 de escala de razón.

1. **Cliente.desde** (v_1) es la fecha de registro del cliente en el sistema.

2. **Edad** (v_2) es el número de años cumplidos a la fecha en la que se extrajo la informacióón.

3. **Sexo** (v_3) es una variable categórica binaria que se codificó con uno en caso de ser mujer y cero en caso de ser hombre.

4. **Ciudad** (v_4) es la población donde el cliente reside.

5. **Código postal** (v_5) es el código postal que pertenece al domicilio qeu registró el cliente.

6. **Colonia** (v_6) es la dirección que el cliente ha indicado como su lugar de residencia; normalmente se toma de la credencial de elector.

7. **Suc** (v_7) es el número de la sucursal en la cual el cliente se registró por primera vez.

8. **Créditos** (v_8) es el número total de créditos que se le han otorgado al cliente.

9. **Vigente** (v_9) número de créditos que se encuentran activos, es decir, que el monto no ha sido saldado y el cliente se encuentra al corriente con los pagos.

10. **Monto.prom** (v_{10}) es la cantidad promedio de dinero otorgada al cliente por cada crédito emitido.

11. **Cred.perd** (v_{11}) es el núúmero de créditos que el cliente ha perdido, es decir, que no ha podido pagar y su prenda ha salido a la venta.

12. **Int.pag** (v_{12}) es el núúmero de créditos en los que se realizaron pagos fuera del esquema de pagos pactado.

13. **Metal** (v_{13}) es el número de prendas metálicas que ha empreñado en total a lo largo del tiempo.

14. **Electrónico** (v_{14}) es el número de prendas de tipo electrónico que ha empreñado el cliente en total a lo largo del tiempo.

15. **Cred.trad** (v_{17}) es el número de veces que el cliente ha escogido esquema tradicional de pagos para liquidar los prestamos.

16. **Cred.PF** (v_{18}) es el número de veces que el cliente ha escogido el esquema de pagos fijos.

17. **Cred.Flex** (v_{19}) es el número de veces que el cliente ha seleccionado el esquema de pagos flexible.

18. **Ing.tot** (v_{20}) es el total de dinero que ha recibido la empresa por los pagos de los préstamos que ha tenido desde que ingresó al sistema.

19. **Ing.prom** (v_{21}) es el ingreso promedio que la empresa ha recibido por los pagos que el cliente ha realizado para cubrir sus préstamos que ha tenido desde que ingresó al sistema.

4.2.2. Análisis Exploratorio

Antes de comenzar a analizar los datos contenidos en la base, se realizó una depuració de la base de datos y esto fue lo que se encontró:

- 119 clientes con errores en las fechas de nacimiento, ya que las fechas fueron mal capturadas desde un inicio en el sistema.
- 761 individuos que presentan datos faltantes en variables como código postal, muncipio, sucursal, monto promedio e ingreso promedio.

Las correlaciones más significativas en esta tabla (4.1) son la de saldo contra monto total con .984, saldo contra ingreso total con .785, monto total contra ingreso total con .837 y saldo contra sucursal con .942.

Aquí en la figura 4.1 se puede preciar como están altamente correlacionadas

Tabla 4.1: Correlaciones

		Suc	Creditos	Vigentes	Saldo	monto_tot	Ing_tot
Suc	Correlación de Pearson	1	.031**	0.01	0	0.002	0.006
	Sig. (bilateral)		0	0.088	0.942	0.722	0.316
	N	30576	30576	30576	30576	30576	30576
Creditos	Correlación de Pearson	.031**	1	.633**	.126**	.212**	.341**
	Sig. (bilateral)	0		0	0	0	0
	N	30576	30576	30576	30576	30576	30576
Vigentes	Correlación de Pearson	0.01	.633**	1	.123**	.131**	.168**
	Sig. (bilateral)	0.088					0
	N	30576	30576	30576	30576	30576	30576
Saldo	Correlación de Pearson	0	.126**	.123**	1	.984**	.785**
	Sig. (bilateral)	0.942					0
	N	30576	30576	30576	30576	30576	30576
monto_tot	Correlación de Pearson	0.002	.212**	.131**	.984**	1	.837**
	Sig. (bilateral)	0.722					0
	N	30576	30576	30576	30576	30576	30576
Ing_tot	Correlación de Pearson	0.006	.341**	.168**	.785**	.837**	1
	Sig. (bilateral)	0.316	0	0	0	0	
	N	30576	30576	30576	30576	30576	30576

**La correlación es significativa al nivel 0.01 (bilateral)

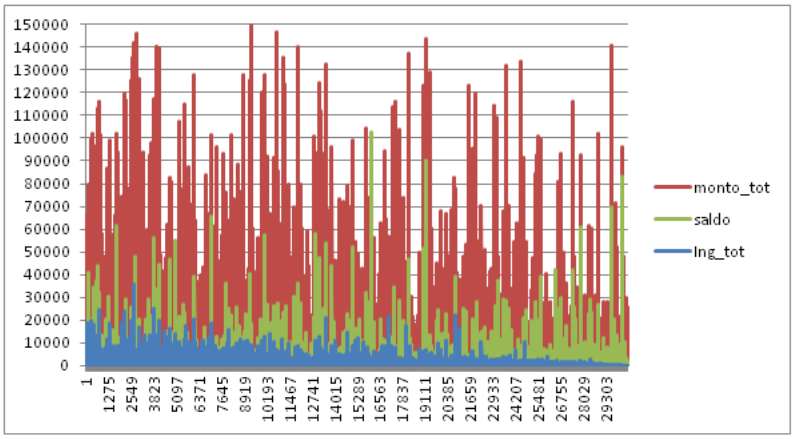


Figura 4.1: Comparación de saldo, monto e ingreso por cliente

las tres variables: saldo, monto e ingreso.

En las gráficas (4.2.2), (4.2.2)y 84.4) se puede observar que tanto el ingreso como el saldo y el monto total por cliente, respectivamente, va disminuyendo conforme el cliente es más reciente.

Acontinuación se muestran como está distribuida la base por sucursales.

En esta gráfica (4.2.2) se puede apreciar como se compara el monto total,

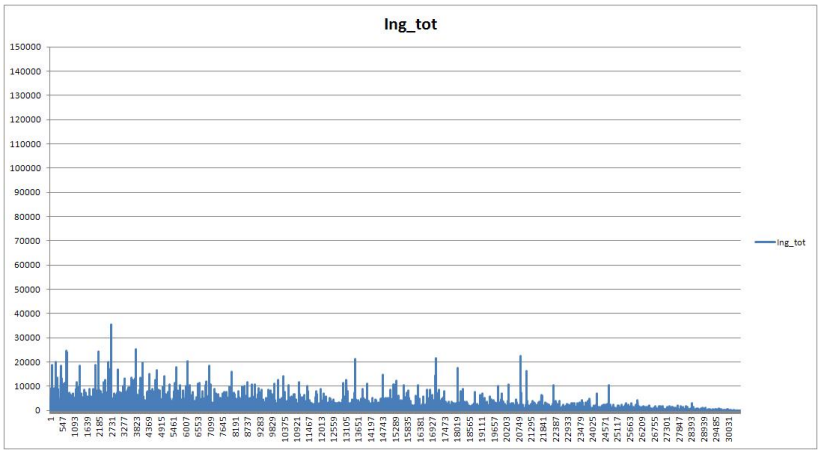


Figura 4.2: Ingreso por cliente

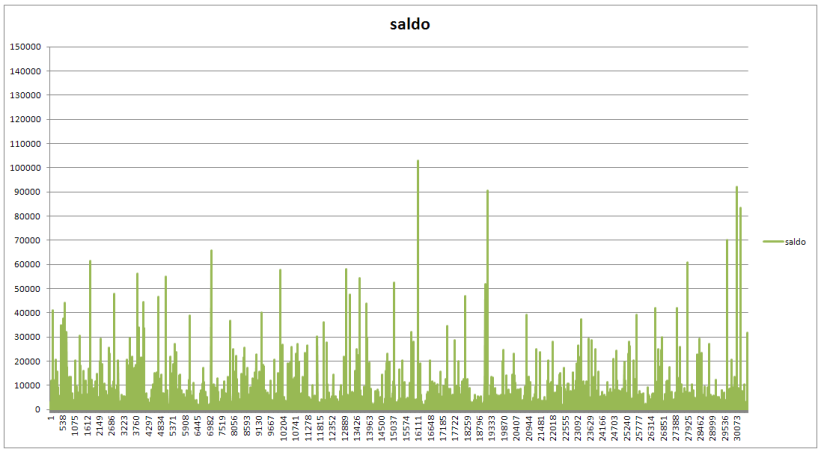


Figura 4.3: Saldo por cliente

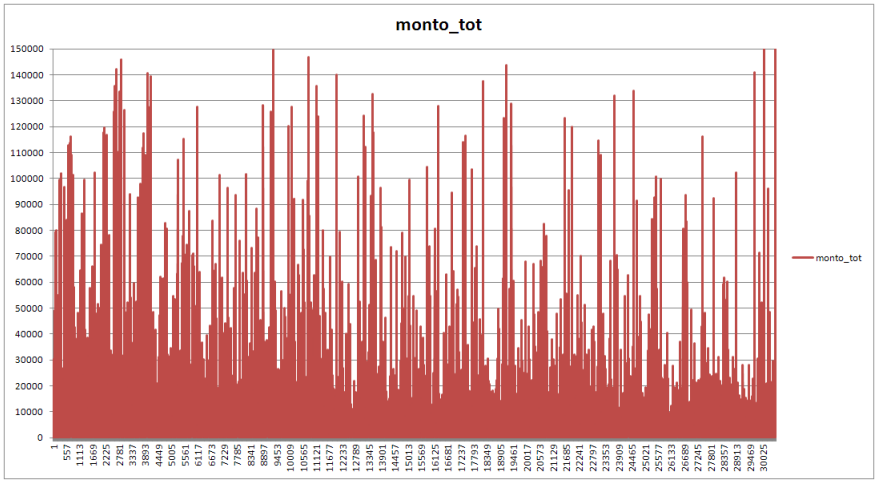
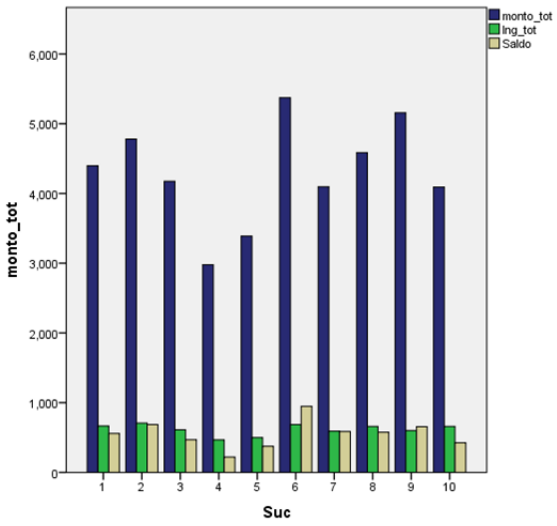


Figura 4.4: Monto total por cliente



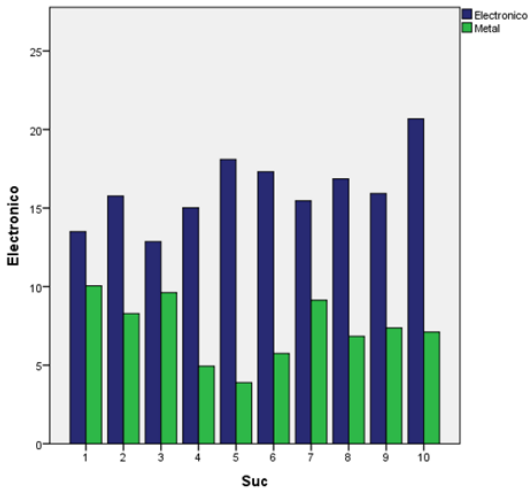


Figura 4.5: Electrónicos por sucursal

ingreso total y saldo por sucursal.

En este caso, en la gráfica (4.2.2), se muestra el empeño de productos electrónicos por sucursal.

En la gráfica (4.2.2) se puede ver el número total de créditos vigentes con los que cuentan las sucursales, a la fecha en la que fue obtenida la información.

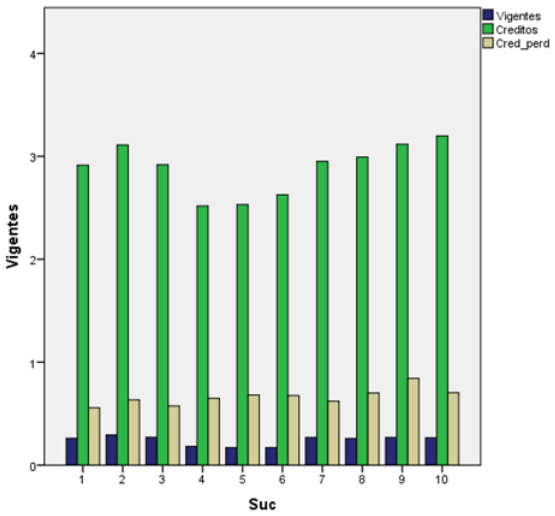


Figura 4.6: Número de créditos vigentes por sucursal

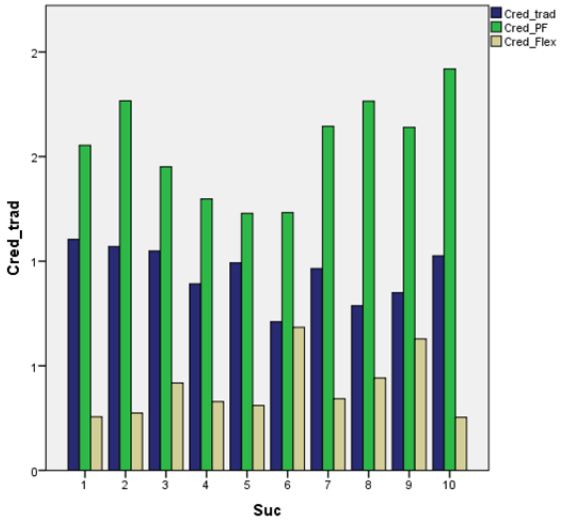


Figura 4.7: Número de créditos otorgados por sucursal

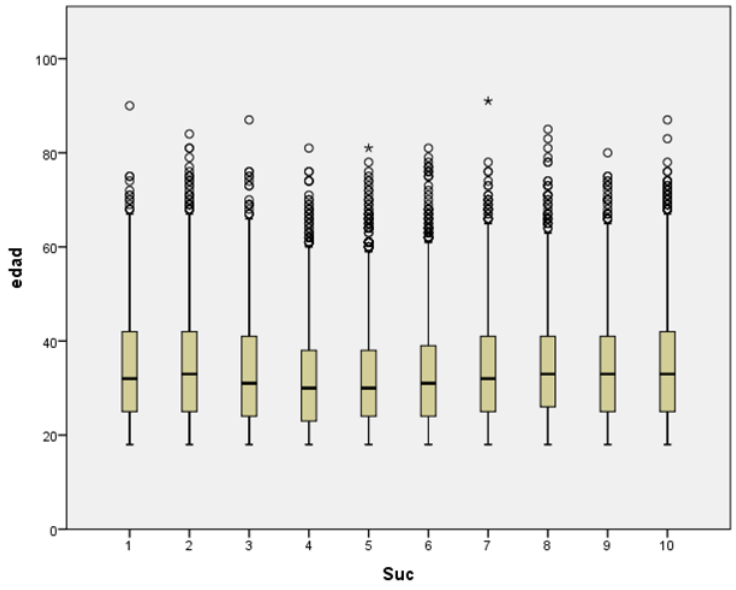


Figura 4.8: Distribución de las edades de los clientes por sucursal

En (4.2.2) se puede ver cuales son las sucursales que han otorgan más créditos.

Ahora, en la gráfica de caja y brazos (4.2.2) muestra como se distribuyen a las edades según la sucursal.

Bibliografía

Bernardo, J. M. (1998). Bruno de finetti en la estadística contemporanea. *Historia de la Matématica en el siglo XX, S. Rios (ed.), Real Academia de Ciencias, Madrid*, 63–80.

Congdon, P. (2007). *Bayesian Statistical Modelling*, Volume 704. John Wiley & Sons.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian Data Analysis*, Volume 2. Chapman and Hall/CRC Boca Raton, FL, USA.

Hall, B. (2012). Bayesian inference. *Statisticat, LLC*.

Marin, J.-M., K. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics 25, 459–507.*

Bibliografía

Martínez Ovando, J. C. (2004). Un criterio predictivo de selección de modelos para series de tiempo.

Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology).