Modelos probabilisticos

Eduardo Tapia
Maestria en Ciencias de la Computación
Cimat
Guanajuato, Mexico
eduardo.tapia@cimat.mx

I. Introducción

En una gran cantidad de aplicaciones, se cuenta con una gran cantidad de datos y características y para poder estudiarlos y obtener información de ellos se suele intentar encontrar la distribución de probabilidad que los genera, ya que en la practica estas no se conocen.

Al no tener una una metodología 100% teórica para realizar estas estimaciones de distribución, se han generado distintos métodos con este objetivo, sin embargo gran cantidad de casos presentan un problema; Muchas veces las características del conjunto de datos con los que se debe trabajar no son independientes entre si y tampoco tienen distribuciones normales, lo que aumenta considerablemente la complejidad del proceso necesario para encontrar una distribución que los pueda modelar.

De aquí nace la necesidad de generar modelos, en los que se pueden presentar relaciones entre las distintas variables de diversas formas resulta importante, ya que nos puede modelar de una mejor manera los datos.

En esta tarea se realizará el diseño de modelos probabilisticos para secuencias bianrias, con el objetivo de generar modelos de tipo independiente, cadena (MIMIC) , arbol de dependencia (Chow-Liu) y todo conectado.

II. Conceptos previos

Para poder realizar modelos probabilísticos, es necesario estar familiarizado con algunos conceptos que son bastante importantes para poder entender el funcionamiento de estos modelos, así como los procesos con los que se construyen.

A. Información Mutua

La información mutua entre 2 variables $I(X,Y)$ es una medida de la dependencia entre dos variables, que nos indica que tanta información podemos obtener de una variable, al observar a la otra. Es decir, cuanto podemos aprender de la variable X únicamente observando Y o viceversa. La información mutua se define como sigue:

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j) log\left(\frac{P(x_i,y_j)}{P(x_i)P(y_j)}\right) \quad (1)$$

B. Entropía de Datos

En la teoría de información, la entropía es la media de la "información" o "Incertidumbre" que contiene un experimento X, de manera que mientras mayor sea la entropía la información dentro del dataset es mayor, mientras que a menor entropía resulta mas sencillo predecir el comportamiento de las variables. La entropía de Shannon se calcula de la siguiente forma

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i)) \quad (2)$$

Donde n es el numero de estados posibles para la variable aleatoria X.

C. Verosimilitud

La verosimilitud o función de verosimilitud, es la medida de ajuste entre los parámetros estimados de la distribución con los datos que se tienen. En la mayoría de los casos, al buscar ajustar una distribución, se busca maximizar la verosimilitud, para que el modelo tenga una gran capacidad de reproducir el dataset.

Dado una muestra $x = \{x_1, x_2, x_3 ... x_n\}$ y una familia de funciones de densidad de probabilidad $f?\theta$ con parámetros θ, la verosimilitud se define:

$$L(\theta) = L(\theta|x) = \prod_{i=1}^{n} f_\theta(x_i) \quad (3)$$

Sin embargo esta representacion de la verosimilitud se encuentra con un problema algebraico, si se presenta algún caso tal que $P(x_i = 0$ todo el producto se desvanece, es por ello que se suele utilizar la Log-verosimilitud en su lugar la cual permite que existan casos con probabilidad 0:

$$log\left(L(\theta)\right) = log\left(L(\theta|x) = \sum_{i=1}^{n} log\left(f_\theta(x_i)\right)\right. \quad (4)$$

D. Divergencia Kullback-Leibler

La divergencia Kullback-Leibler es una medida de similitud entre 2 distribuciones de probabilidad P, Q, Está definida como:

$$D_{KL}(P||Q) = \sum_{x\in\mathcal{X}} P(x) log\left(\frac{P(x)}{Q(x)}\right) \quad (5)$$

Donde P y Q viven en el mismo espacio de probabilidad y P es la distribución de referencia, si esta distancia es 0, significa que las distribuciones son iguales.

III. Modelos a utilizar

Existen diversas clases de modelos sobre las que se pueden trabajar, a continuación se presentarán los que se utilizarán en esta tarea. Los modelos serán

1) Modelo Independiente
2) Modelo MIMIC
3) Modelo Cow-Liu
4) Modelo Todo conectado.

A. Modelo independiente

El modelo independiente es el modelo mas simple que se puede crear, ya que dado que se considera que las variables son independientes entre si, simplemente se tiene que calcular la probabilidad de cada variable aleatoria en si misma, y a partir de ello se genera el modelo, para esta clase de modelos, el grafo que los representa es un grafo de puros nodos sin interconexiones, donde la probabilidad de cada nodo corresponde a $P(x_i = 1)$.

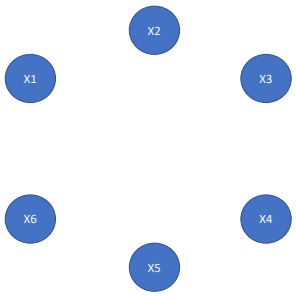

Fig. 1. Modelo Independiente

Con funcion de probabilidad

$$P(X) = P(x_1)P(x_2)P(x_3)P(x_4)P(x_5)P(x_6) \quad (6)$$

B. Modelo de cadena

El modelo de cadena, tambien conocido como MIMIC[1], es un modelo que busca modelar el comportamiento de los datos mediante las relaciones entre las variables aleatorias, generando modelos en los cuales todas las variables a excepción de la raíz, dependen de otra variable, este modelo tiene 2 características principales:

1) Cada nodo solo puede tener un padre
2) Cada nodo solo puede tener un hijo

Esto genera un grafo que tiene una estructura muy similar a la que se puede observar en la figura 2
Este m modelo tiene tiene la siguiente función de probabilidad

$$P(X) = P(x_3)P(x_1|x_2)P(x_2|x_3,x_1)P(x_6|x_3,x_1,x_2) * P(x_5|x_3,x_1,x_2,x_6)$$

[1]Minimum Mutual Information for Imput Clustering

Fig. 2. Grafo MIMIC

La metodologia para armar un grafo MIMIC esta descrita en el algoritmo 1

Algorithm 1 Algoritmo MIMIC

1: Se calcula la entropía de las variables
2: Se calcula la tabla de Información mutua entre todas las variables.
3: Se toman las 2 variables con MAYOR información mutua y se agregan al modelo M considerándolos como nodos externos
4: Se crea una lista con las variables restantes L
5: **while** $L! = \{\varnothing\}$ **do**
6: **for** i en $Nodos\_externos$ de M **do**
7: Buscar en la lista L la variable con mayor información mutua respecto al nodo externo
8: Agregar el nodo a M y conectarlo al nodo externo correspondiente y actualizar este nuevo nodo como nodo externo y retirar el anterior.
9: Retirar el nodo de la lista L
10: **end for**
11: **end while**
12: Regresar M

C. Árbol de Dependencias

El árbol de dependencias es un modelo propuesto por Chow-Liu [Chow and Liu, 1968] en el cual proponen un metodo distinto al MIMIC cambiando la forma, de una cadena que solo permite un padre y un hijo a un árbol de probabilidad.

La principal ventaja que tienen estos modelos respecto a los modelos de cadena o el independiente, es que busca realizar las conexiones entre las variables de manera tal que se maximice la información mutua dentro de todo el modelo, manteniendo las relaciones de mayor interés dentro de las variables a modelar. Las características que definen sencillamente este modelo son:

1) Cada nodo puede tener solo un padre
2) Cada nodo puede tener cualquier numero de hijos

Esto permite mejorar la calidad del modelo, y da como resultado un grafo similar al de la figura 3

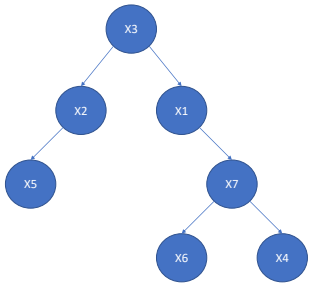

Fig. 3. Grafo para el modelo Chow-Liu

Algorithm 2 Algoritmo Chow-Liu

1: Se calcula la tabla de Información mutua entre todas las variables.
2: Se toman las 2 variables con MAYOR información mutua y se agregan al modelo M
3: Se crea una lista con las variables restantes L
4: **while** $L! = \{\varnothing\}$ **do**
5: **for** $i = nodo$ en L **do**
6: Buscar en la tabla de información mutua el nodo que tenga mayor IM con alguno de los nodos que ya estén en M
7: Conectar ese nodo con el que corresponda en M.
8: Retirar el nodo de la lista L
9: **end for**
10: **end while**
11: Regresar M

D. Modelo Todo conectado

El modelo todo conectado corresponde aquel que conecta todos los nodos con todos, esto genera un modelo donde todas las variables están condicionadas con todas, de manera que conforme se va calculando una variable, la siguiente depende de todas las anteriores.

La metodología para construir un modelo Todo conectado se puede observar en el algoritmo 3

Algorithm 3 Algoritmo Todo conectado

1: Se calcula la tabla de probabilidades condicionales de las variables.
2: Se toma la variable raíz y esta se conecta con todas las demás variables con mayor información mutua y se agregan al modelo M considerándolos como nodos externos
3: Regresar M

Este modelo genera un grafo con una topología similar a la del grafo 4

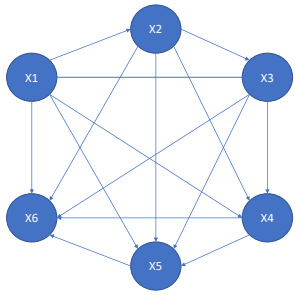

Fig. 4. Grafo Todo conectado

Y su funcion de probabilidad es:

$$P(X) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2) * P(x_4|x_1,x_2,x_3)P(x_5|x_1,x_2,x_3,x_4) * P(x_6|x_1,x_2,x_3,x_4,x_5)$$

IV. Metodología

Para realizar los modelos es necesario contar con un conjunto de datos, para ello se realizaron 2 datasets, cada uno con su propia regla de creación, las cuales se pueden ver en las tablas I y II

Probabilidades BD1
$P(x_1 = 1) = 0.2$
$P(x_2 = 1\|x_1 = 0) = 0.92$
$P(x_2 = 1\|x_1 = 1) = 0.31$
$P(x_3 = 1\|x_2 = 0) = 0.42$
$P(x_3 = 1\|x_2 = 1) = 0.15$
$P(x_4 = 1\|x_3 = 1) = 0.75$
$P(x_5 = 1\|x_4 = 0) = 0.67$
$P(x_5 = 1\|x_4 = 1) = 0.43$
$P(x_6 = 1\|x_5 = 0) = 0.35$
$P(x_6 = 1\|x_5 = 1) = 0.82$

TABLE I
TABLA DE PROBABILIDADES PARA BASE DE DATOS 1

Probabilidades BD2
$P(x_1 = 1) = 0.2$
$P(x_2 = 1\|x_1 = 0) = 0.92$
$P(x_2 = 1\|x_1 = 1) = 0.31$
$P(x_4 = 1\|x_1 = 0) = 0.42$
$P(x_4 = 1\|x_1 = 1) = 0.15$
$P(x_5 = 1\|x_2 = 0) = 0.28$
$P(x_5 = 1\|x_2 = 1) = 0.75$
$P(x_6 = 1\|x_2 = 0) = 0.67$
$P(x_6 = 1\|x_2 = 1) = 0.43$
$P(x_3 = 1\|x_6 = 0) = 0.35$
$P(x_3 = 1\|x_6 = 1) = 0.82$

TABLE II
TABLA DE PROBABILIDADES PARA BASE DE DATOS 2

V. Resultados

A. Base de datos 1

Partiendo de la tabla I se genero una base de datos de 60000 muestras, y a partir de ellos se generaron los modelos previamente mencionados, la verosimilitud y divergencia K-L se presentan en la tabla III:

Modelo	Log-Likelihood	Divergencia K-L
Independiente	-2.111448e+05	-1.270861e+03
Cadena	-1.863984e+05	8.529132e-02
Dependence Tree	-1.863984e+05	8.529132e-02
Todo conectado	-9.125040e+05	-5.425638e-01

TABLE III
Resultados para la base de datos 1

B. Base de datos 2

Para la segunda base de datos basada en la tabla II se generó un conjunto de 60000 muestras, y se realizaron las mismas medidas las cuales se presentan en la tabla. IV

Modelo	Log-Likelihood	Divergencia K-L
Independiente	-2.203699e+05	-6.099957e+02
Cadena	-1.976322e+05	-8.986485e-01
Dependence Tree	-1.967206e+05	7.994885e+00
Todo conectado	-8.810738e+05	-4.370360e-01

TABLE IV
Resultados para la base de datos 2

C. Resultados para las cadenas aleatorias

1) Db1: Se eligieron 3 cadenas aleatorias y se calculo la probabilidad de que se generasen con cada uno de los modelos, las cadenas que se eligieron del primer dataset fueron:

- 010011
- 010111
- 100011

Y las probabilidades que se calcularon son las siguientes:

Cadena	$P(x\|T_1)$	$P(x\|IND)$	$P(x\|TC)$	$P(x\|MIMIC)$	$P(x\|TREE)$
010011	2.474673e-01	1.136828e-01	5.586043e-08	2.473101e-01	2.473101e-01
010111	6.176424e-02	6.861804e-02	3.138843e-08	6.180721e-02	6.180721e-02
100011	3.166126e-02	7.232291e-03	5.718283e-08	3.146416e-02	3.146416e-02

TABLE V
Probabilidades cadenas db1

2) Db2: Para la segunda tabla se eligieron las siguientes cadenas:

- 101000
- 010110
- 011000

Cadena	$P(x\|T_1)$	$P(x\|IND)$	$P(x\|TC)$	$P(x\|MIMIC)$	$P(x\|TREE)$
101000	9.754668e-03	2.550213e-03	8.203921e-07	1.348761e-02	9.551285e-03
010110	8.589672e-02	3.446946e-02	1.119834e-07	8.323232e-02	8.626366e-02
011000	2.129064e-02	4.165493e-02	7.935091e-07	1.741540e-02	2.109171e-02

TABLE VI
Probabilidades cadenas db2

VI. Conclusiones

En las tablas III y IV se puede observar que los modelos con mayor verosimilitud es el todo conectado, lo cual tiene bastante sentido ya que es el modelo que contiene todas las relaciones posibles entre las distintas variables aleatorias, sin embargo con modelos muy grandes esto se puede llegar a volver pesado al momento de realizar las operaciones.

Por otro lado podemos observar que aun cuando el modelo todo conectado e independiente tienen una mayor verosimilitud que los modelos MIMIC y el Cho-liu, estos ultimos resultan mejor opción al momento de reproducir la distribución, ya que en ambos casos, la Divergencia de Kullback-Leibler resulta ser mucho menor que en los otros casos, esto es relevante ya que el realizar evaluaciones de estos modelos es mucho mas rápido considerando el numero necesario de operaciones, en funcion de las variables aleatorias a reproducir y lo que se quiere es tener un modelo que reproduzca las propiedades de la distribución y que al mismo tiempo maximice la verosimilitud, de manera que para el caso de estos modelos, es mas recomendable implementar el MIMMIC o Cho-Liu en lugar de un todo conectado o un independiente.

References

[Chow and Liu, 1968] Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.