UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

ANÁLISIS DE DATOS

LABORATORIO 4: CLASIFICADOR BAYESIANO

Autores:
Nicolás Mariángel Toledo
Juan Pablo Rojas Rojas
Profesor:
Max Chacón Pacheco
Ayudante:
Ignacio Ibáñez Aliaga

Santiago - Chile
25 de diciembre de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

El cuerpo humano está conformado por distintos sistemas que hacen de sí un ente completo y complejo con un organismo en equilibrio. Uno de estos sistemas es el *Sistema Endocrino* que tiene la función principal de regular la producción de hormonas, las cuales se encargan de influir en la mayoría de las funciones del organismo, por ejemplo: funcionamiento de órganos, control de las distintas funciones del organismo, auto-regulación, comportamientos del individuo, control de homeostasis, entre otras funciones. Así el sistema endocrino posee órganos que ayudan a formar el sistema como tal, uno es estos órganos es la llamada *Tiroides*, la cual es la encargada de producir hormonas tiroideas que ayudan al cuerpo utilizar energía, mantener la temperatura corporal y a que el cerebro, el corazón, los músculos y otros órganos funcionen normalmente [1].

No siempre nuestros sistemas y órganos funcionan como corresponde, en el caso de la Tiroides uno de los problemas más comunes son el *Hipotiroidismo* e *Hipertiroidismo* que tiene que ver principalmente con el desequilibrio de producción de *hormonas tiroideas*, en esta oportunidad el tema de interés se centra en el Hipotiroidismo, que se ocurre cuando la tiroides no es capaz de producir suficientes hormonas tiroideas T3 y T4. Esto último es un tema con gran potencial de estudio, ya que la población actual gradualmente ha incrementado este tipo de problemas, por lo cual se busca tener medicina suficientemente apta para dar soluciones eficientes a estos problemas, para esto se tiene evidencia de distintos datos de pacientes que nos pueden permitir determinar patrones de comportamientos y causas de tal enfermedad, lo que puede concluir en una gran aporte a la medicina para determinar causas de estos fallos en el organismo.

1.2 OBJETIVOS

- Definir la clase a determinar con el Clasificador Bayesiano ingenuo.
- Determinar los atributos utilizados y conjuntos de entrenamiento y de test.
- Analizar el clasificador final obtenido y comparar con experiencias anteriores.

1.3 ORGANIZACIÓN DEL DOCUMENTO

El documento consta de 4 capítulos principales: Un marco teórico donde se explican conceptos de importancia para la experiencia, la descripción del proceso de obtención del Clasificador Bayesiano con el respectivo análisis, y por último finalizar con las conclusiones respecto a la experiencia.

CAPÍTULO 2. MARCO TEÓRICO

2.1 CLASIFICADOR BAYESIANO INGENUO

Bayesiano ingenuo es un algoritmo muy utilizado para resolver problemas de clasificación. El modelo se denomina naïve porque trata todas las variables de predicción propuestas como independientes unas de otras. El bayesiano ingenuo es un algoritmo rápido y escalable que calcula las probabilidades condicionales para las combinaciones de atributos y el atributo de objetivo. A partir de los datos de entrenamiento se establece una probabilidad independiente. Esta probabilidad proporciona la verosimilitud de cada clase objetivo, una vez dada la instancia de cada categoría de valor a partir de cada variable de entrada [2].

2.2 PROBABILIDAD A PRIORI

Sea una clase definida como c_i, la probabilidad a priori $p(c_i)$ se define como la probabilidad de que un sujeto clasifique en dicha clase c_i:

$$p(c_i) = \lim_{n \to +\infty} \frac{n_i}{n} \tag{2.1}$$

2.3 PROBABILIDAD A POSTERIORI

Indica la probabilidad de que un sujeto pertenezca a una clase c_i dada una condición x. Tener el valor de x dependerá del hecho posterior de que la variable de características x sea medida[3]. Para dos clases se tiene:

$$\sum_{i=1}^{2} p(c_i/x) = 1 \tag{2.2}$$

CAPÍTULO 3. OBTENCIÓN DEL CLASIFICADOR

3.1 PRE- PROCESAMIENTO

El trabajo con el dataset implica manejar una gran cantidad de datos, es por esto que es necesario analizar la conformación de datos de modo que la manipulación en procedimientos futuros sea más consistente. Para esto se debe analizar registros perdidos, valores atípicos o datos que no aporten información relevante para el estudio del problema, de forma que las conclusiones no se vean afectadas por anomalías de la base de datos.

3.1.1 Eliminación de registros nulos

El dataset utilizado de hipotiroidismo posee 2800 sujetos originalmente, existiendo varios sujetos que tienen registros nulos (marcados con '?' en esta base de datos), es decir, hubo varios sujetos los cuales no se realizaron alguno de los exámenes. La cantidad de sujetos con datos incompletos se muestra especificada en la siguiente tabla:

Tabla 3.1: Detalles de variables con datos incompletos.

Variable	Porcentaje '?' [%]	Número de incidencias
sex	3.928571	110
TSH	10.14286	284
T3	20.89286	585
TT4	6.571429	184
T4U	10.60714	297
FTI	10.53571	295
TBG	100	2800

Se decide en contra de la imputación, dado que esta asume que los datos son aleatorios, pero en este caso es más probable que el doctor decida realizar todos los exámenes sólo cuando el sujeto presente síntomas de hipotiroidismo, en concordancia a los resultados obtenidos parcialmente por algunos de los exámenes. Si el doctor decide lo contrario, es decir, no someter al sujeto a todos los exámenes, esto podría significar que el sujeto se encuentra saludable (o que padece de otra enfermedad no relacionada con la tiroides), efecto que sería ocultado por la imputación de datos. Por estas razones se decide eliminar los pacientes que tienen registros incompletos.

3.1.2 Eliminación de variables

Se decide eliminar ciertas variables de acuerdo con los siguientes criterios:

- Variables que no contengan ningún dato o que todos sus datos sean registros desconocidos o nulos (NA o '?').
- Que tal variable no entregue información útil para el estudio que se quiere realizar.

3.1.2.1 TBG

Esta variable corresponde a el resultado de una medición de TBG, el cuál es un examen para medir el nivel de globulina fijadora de tiroxina, glucoproteína que lleva hormona tiroidea a través de la sangre. Ninguno de los pacientes fue sometido a este examen en la base de datos, razón suficiente para eliminar esta variable del análisis, dado que todos los datos de esta variable son '?'. Se eliminan las variables *TBG* y *TBG measured* de la base de datos.

3.1.2.2 Fuente de referencia

Esta variable indica las referencias que ocupa el autor de la base de datos, es decir, la persona que recopiló toda la información que tiene esta base de datos. Para el agrupamiento de datos, esta variable no entrega información útil, todo lo contrario, podría crear agrupaciones según donde se obtuvieron los datos, lo que sería contraproducente para el análisis de los datos. Se elimina la variable *referral.source* de la base de datos.

3.1.2.3 Variables de medición

Son variables que definen si un cierto sujeto fue sometido o no a un examen. Como se mencionó anteriormente, se eliminan todos los sujetos que no fueron sometidos a todos los exámenes que presenta la base de datos, por consecuencia de esto, las variables de medición ahora presentan todas un solo valor que indica que el sujeto se realizó el examen ('t'), por lo que esta variable ya no entrega información útil para cualquier tipo de análisis. Las variables de medición a eliminar son las siguientes:

- **TSH measured:** Indica si el paciente ha sido sometido a pruebas para medir TSH.
- **T3 measured:** Indica si el paciente ha sido sometido a pruebas para medir la hormona triiodotiroxina (T3).
- **TT4 measured:** Indica si el paciente ha sido sometido a mediciones de T4.
- **T4U measured:** Indica si el paciente ha sido sometido a pruebas de T4U.
- **FTI measured:** Indica si el paciente ha sido sometido a mediciones de FTI.

3.2 CLASE

Luego del procedimiento descrito en la sección anterior, para realizar los procesos de clasificación es importante definir la clase que se utilizará para realizar dicha clasificación. El problema que se ha abarcado en desde la primera experiencia es el caso del hipotiroidismo, en definitiva para este caso lo que se quiere conseguir es saber si una persona padece hipotiroidismo en base a una probabilidad que a partir de exámenes hormonales, edad, sexo y otros factores se puede clasificar dentro de dicha anomalía.

Al analizar la base de datos, la variable que mayor importancia tiene en la clasificación de hipotiroidismo es la variable de nombre *clasification*, por ende se denomina como atributo predictor ya que es capaz de definir con mejor aproximación si un paciente clasifica dentro de esta enfermedad. Cabe destacar que dado que la enfermedad posee varios síntomas, no existe un único atributo capaz de definir si un paciente tiene hipotiroidismo los cuales claramente se tendrán en consideración para realizar el clasificador bayesiano.

Abordando el Clasificador Bayesiano, se aplica clasificador Bayesiano ingenuo basado en la premisa que los atributos considerados para su construcción son independientes entre sí basandose en información previa o evidencia. A partir de loa anterior se debe definir una base de datos de entrenamiento, para este caso *allhypo.data* y una base de datos de pruebas, para este caso *allhypo.test* ambas apuntadas al estudio del hipotiroidismo, al realizar este procedimiento se asume con los datos son representativos ya que en ambas bases de datos se tiene datos con igual atributos.

3.3 CONJUNTOS UTILIZADOS

Las variables que se utilizan con un total de 22 son los siguientes:

1. age (variable continua)
2. sex (variable categorica)
3. on thyroxine (variable categorica)
4. query on thyroxine (variable categorica)
5. on antithyroid medication (variable categorica)
6. sick (variable categorica)
7. pregnant (variable categorica)
8. thyroid surgery (variable categorica)
9. I131 treatment (variable categorica)
10. query hypothyroid (variable categorica)
11. query hyperthyroid (variable categorica)
12. lithium (variable categorica)

13. goitre (variable categorica)
14. tumor (variable categorica)
15. hypopituitary (variable categorica)
16. psych (variable categorica)
17. TSH (variable continua)
18. T3 (variable continua)
19. TT4 (variable continua)
20. T4U (variable continua)
21. FTI (variable continua)
22. results (ATRIBUTO PREDICTOR)

Luego de tener una base de datos de entrenamiento limpia, el total de datos es de 1946. Por otra parte, la base de datos de prueba tiene un total de 696 datos.

CAPÍTULO 4. ANÁLISIS DE RESULTADOS

4.1 CLASIFICACIÓN

Luego de entrenar el clasificador con la base de datos de entrenamiento descrita anteriormente, los resultados obtenidos se pueden presentar en una matriz de confusión, comparando las predicciones de las clases con sus instancias reales correspondientes a la base de datos de prueba. La siguiente tabla presenta la matriz de confusión del clasificador bayesiano ingenuo donde las filas corresponden a las instancias obtenidas con la clasificación y las columnas a los datos obtenidos de la base de datos de prueba.

Tabla 4.1: Matriz de confusión del clasificador bayesiano ingenuo.

Resultados	**compensated hypothyroid**	**negative**	**primary hypothyroid**
compensated hypothyroid	21	3	1
negative	7	611	0
primary hypothyroid	1	3	19
secondary hypothyroid	0	0	0

Cabe destacar que la clase ”secondary hypothyroid”solo está presente en la clase obtenida desde el clasificador y no en la real, esto debido a que esta clase solo se contempla en los datos de entrenamiento no así en los datos de prueba. Así se presenta dicha fila correspondiente a la instancia de la clase ausente de datos.

La matriz de confusión permite obtener medidas de rendimiento respecto del clasificador con los datos de prueba, se pueden encontrar índices para medir precisión general, como también por clase como también valores predictivos. En la siguiente tabla se presentan los índices de precisión calculados a partir de la matriz de confusión.

Tabla 4.2: Índices de precisión del clasificador.

Medida	**Proporción**	**Porcentaje**
Precisión total	0.977	97,7 %
Valor predictivo: negative	0.989	98,9 %
Valor predictivo: primary hypothyroid	0.826	82,6 %
Valor predictivo: compensated hypothyroid	0.84	84 %

A partir de lo anterior se puede asegurar la precisión del clasificador debido a que este clasificó con un 97.7 % cometiendo así un error del 2.3 % por otro lado las la predicción del clasificador de acuerdo a cada una de las clases: negative, primary hypothyroid y compensated hypothyroid, cometió un error de 1.1 %, 17.4 % y 16 % respectivamente.

4.2 COMPARACIÓN CON EXPERIENCIAS ANTERIORES

Anteriormente se ha trabajado con distintos métodos: método de agrupamiento k-medias y reglas de asociación, es por esto que es factible realizar comparaciones de acuerdo a lo experimentado.

4.2.1 k-medias

Para partir cabe destacar que k-medias es un método de aprendizaje no supervisado el cual actúa en base al agrupamiento de las observaciones utilizando como base la distancia entre ellas, por otro lado el clasificador bayesiano ingenuo es un método de aprendizaje supervisado con probabilidades a priori y posteriori generando así una clasificación basada en la máxima probabilidad que tiene una observación para clasificar en cierta clase. Dicho lo anterior claramente no se puede realizar una comparación certera por ser métodos de aprendizaje totalmente distintos. Aún así en el desarrollo de la experiencia de k-medias, los sujetos se clasifican en cluster debido en base a los posibles factores que podrían causar el padecimiento de hipotiroidismo, entre estos esta el sexo, y las diversas hormonas que están presente en el problema, que por otro lado en la experiencia actual los procedimiento concluyen en el tipo del cuadro de hipotiroidismo del paciente logrando determinar la clasificación además de evaluar de su precisión.

4.2.2 Reglas de asociación

A diferencia del caso anterior, en esta oportunidad ambos métodos son de aprendizaje supervisado, los cuales tienen como objetivo predecir el cuadro del paciente según las distintas variables, la comparación se realiza en cuanto a su metodología. Las reglas de asociación como se mencionó en el marco teórico buscan automáticamente las asociaciones que se podrían encontrar manualmente usando técnicas de visualización, así obtenido el conjunto de reglas de asociación por el que se caracteriza permiten tener como consecuencia la clase a la cual pertenece cierto paciente. Por otro lado el clasificador bayesiano ingenuo, a través, de las probabilidades optimiza la probabilidad a posteriori pero con un proceso de caja negra, no pudiendo determinar así las variables que tienen mayor importancia.

Si bien ambos métodos son distintos, mediante reglas de asociación se determinan aquellas reglas que tienen como consecuencia hipotiroidismo 0in especificar el tipo pero si hallando las variables de importancia para el modelo.

CAPÍTULO 5. CONCLUSIONES

La experiencia de este laboratorio permitió aplicar un método basado en las probabilidades a priori y posteriori que al igual que en la experiencia anterior recaen en un método de aprendizaje supervisado.

El preprocesamiento para este tipo de método fue muy importante para clasificar correctamente las clases, esto debido a que generalmente estos datos que no son considerados implican ruido y evitan resultados precisos.

Respecto a los resultados obtenidos la precisión general alcanzó un 97.7 %, que claramente representa un buen aprendizaje del sistema en cuanto a los datos de entrenamiento proporcionados y luego evaluándose con los datos de prueba. El clasificador bayesiano, permitió a diferencia de otros métodos la capacidad de clasificar en distintos tipos de hipotiroidismo, lo cual si bien no se presentan las variables que causan más riesgo de hipotiroidismo, si se puede dar un uso de resultado concluyente en base a un método robusto basado en la evidencia y probabilidad.

Si bien se puede considerar el clasificador bayesiano como un método más preciso y practico que los métodos estudiados anteriormente, esto es cierto en cuanto a método de aprendizaje no supervisados, que por otro lado si entramos en comparaciones con métodos de aprendizaje supervisado a fin de cuentas cada uno posee ventajas y desventajas que pueden favorecer en ciertas situaciones de uso, como lo es el caso de las reglas de asociación.

Luego, abordando el caso del clasificador bayesiano, asumiendo que las variables son totalmente independientes, se pudieron obtener resultados esperados escogiendo aquellos atributos predictores de manera efectiva y prácticamente su implementación en código es bastante sencilla, ahora el desafío y siguiente desafío es aprender un método de clasificación para variables dependientes entre si.

CAPÍTULO 6. BIBLIOGRAFÍA

[1] A. T. Association, *Pruebas De Función Tiroidea*, 2016. dirección: https://www.thyroid.org/las-pruebas-de-funcion-tiroidea/.

[2] D. M. Chacón, *Clasificación Bayesiana*, 2017. dirección: http://www.udesantiagovirtual.cl/moodle2/pluginfile.php?file=%2F217599%2Fmod_resource%2Fcontent%2F1%2FCap%C3%ADtulo%20VI%20An%C3%A1lisis%20de%20Datos_CB_N.pdf.

[3] I. I. K. Center, *Bayesiano ingenuo de Oracle*, 2014. dirección: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/oracle_bayes.html.

CAPÍTULO 7. ANEXO 1 - CÓDIGO R

```
#_____________Universidad de Santiago de Chile_____________________
#________Departamento de Ingenieria en Informatica_________________
#____________________Analisis de Datos_________________________________
#
# Laboratorio 4: Clasificador Bayesiano
# Integrantes: Nicolas Mariangel | Juan Pablo Rojas
# Profesor: Max Chacon
# Ayudante: Ignacio Ibanez Aliaga
#
# http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

# Naive Bayes Classifier
library("e1071")

library(readr)
library(cluster)
library(ggplot2)

preprocessing <- function(allhypo){
  # Limpieza de los resultados (no se usa el ".|numero" solo interesa la clase)
  allhypo$results <- vapply(strsplit(allhypo$results,"\\."), `[`, 1, FUN.VALUE=
      character(1))

  # Se eliminan estas columnas porque no fue medida la TBG en la base de datos.
  allhypo$TBG <- NULL
  allhypo$`TBG measured` <- NULL

  # Luego se procede a cambiar el caracter '?' por NA para poder ejecutar
      complete.cases()
  allhypo[allhypo=="?"] <- NA
  allhypo$age[allhypo$age >= 123] <- NA

  # Se eliminan los datos que tienen NA (originalmente '?')
  allhypo <- allhypo[complete.cases(allhypo), ]
```

```
33
34  # Luego como todas las columnas de si es que se mide o no algo son siempre 't
        '
35  # por haber eliminado los sujetos, no se usarian estas columnas en el
        analisis
36  allhypo$`T3 measured` <- NULL
37  allhypo$`T4U measured` <- NULL
38  allhypo$`TSH measured` <- NULL
39  allhypo$`TT4 measured` <- NULL
40  allhypo$`FTI measured` <- NULL
41
42  # Eliminar referral.source debido a que no aporta informacion
43  allhypo$`referral source` <- NULL
44
45  # Transformar todas las columnas numericas a variables numericas.
46  allhypo <- transform(allhypo, age = as.numeric(age))
47  allhypo <- transform(allhypo, TSH = as.numeric(TSH))
48  allhypo <- transform(allhypo, T3 = as.numeric(T3))
49  allhypo <- transform(allhypo, TT4 = as.numeric(TT4))
50  allhypo <- transform(allhypo, T4U = as.numeric(T4U))
51  allhypo <- transform(allhypo, FTI = as.numeric(FTI))
52
53  # Transformar todas las columnas categoricas a factores
54  allhypo <- transform(allhypo, sex = as.factor(sex))
55  allhypo <- transform(allhypo, on.thyroxine = as.factor(on.thyroxine))
56  allhypo <- transform(allhypo, query.on.thyroxine = as.factor(
        query.on.thyroxine))
57  allhypo <- transform(allhypo, on.antithyroid.medication = as.factor(
        on.antithyroid.medication))
58  allhypo <- transform(allhypo, sick = as.factor(sick))
59  allhypo <- transform(allhypo, pregnant = as.factor(pregnant))
60  allhypo <- transform(allhypo, thyroid.surgery = as.factor(thyroid.surgery))
61  allhypo <- transform(allhypo, I131.treatment = as.factor(I131.treatment))
62  allhypo <- transform(allhypo, query.hypothyroid = as.factor(query.hypothyroid
        ))
63  allhypo <- transform(allhypo, query.hyperthyroid = as.factor(
        query.hyperthyroid))
```

```
64    allhypo <- transform(allhypo, lithium = as.factor(lithium))
65    allhypo <- transform(allhypo, goitre = as.factor(goitre))
66    allhypo <- transform(allhypo, tumor = as.factor(tumor))
67    allhypo <- transform(allhypo, hypopituitary = as.factor(hypopituitary))
68    allhypo <- transform(allhypo, psych = as.factor(psych))
69    allhypo <- transform(allhypo, results = as.factor(results))
70
71    # Eliminar los valores atipicos
72    sdRange <- 3
73    MaxTSH <- mean(allhypo$TSH) + sdRange * sd(allhypo$TSH)
74    MaxT3 <- mean(allhypo$T3) + sdRange * sd(allhypo$T3)
75    MaxTT4 <- mean(allhypo$TT4) + sdRange * sd(allhypo$TT4)
76    MaxT4U <- mean(allhypo$T4U) + sdRange * sd(allhypo$T4U)
77    MaxFTI <- mean(allhypo$FTI) + sdRange * sd(allhypo$FTI)
78    allhypo <- subset(allhypo , (age <= 120) & (TSH <= MaxTSH) & (T3 <= MaxT3) &
          (TT4 <= MaxTT4) & (T4U <= MaxT4U) & ( FTI <= MaxFTI) )
79
80    return(allhypo)
81  }
82
83  # Lectura de los datos de entrenamiento y de prueba + asignarle nombres a las
        columnas de acuerdo a lo escrito en allhypo.names
84  allhypo_training <- read_csv("allhypo/allhypo.data", col_names = FALSE)
85  allhypo_training_Rownames <- c("age", "sex", "on thyroxine", "query on
        thyroxine", "on antithyroid medication", "sick", "pregnant", "thyroid
        surgery", "I131 treatment", "query hypothyroid", "query hyperthyroid", "
        lithium", "goitre", "tumor", "hypopituitary", "psych", "TSH measured", "TSH
        ", "T3 measured", "T3", "TT4 measured", "TT4", "T4U measured", "T4U", "FTI
        measured", "FTI", "TBG measured", "TBG", "referral source", "results")
86  colnames(allhypo_training) <- allhypo_training_Rownames
87
88  allhypo_test <- read_csv("allhypo/allhypo.test", col_names = FALSE)
89  allhypo_test_Rownames <- c("age", "sex", "on thyroxine", "query on thyroxine",
        "on antithyroid medication", "sick", "pregnant", "thyroid surgery", "I131
        treatment", "query hypothyroid", "query hyperthyroid", "lithium", "goitre",
         "tumor", "hypopituitary", "psych", "TSH measured", "TSH", "T3 measured", "
        T3", "TT4 measured", "TT4", "T4U measured", "T4U", "FTI measured", "FTI", "
```

```
    TBG measured", "TBG", "referral source", "results")
colnames(allhypo_test) <- allhypo_test_Rownames

# Bases de datos de entrenamiento y prueba procesadas.
data_training <- preprocessing(allhypo_training)
data_test <- preprocessing(allhypo_test)

# Entrenaminto
model <- naiveBayes(results ~., data = data_training)

# Predecir los datos
results <- predict(object = model, newdata=data_test, type = "class")

# Matriz de confusion - Predecidos vs Entrenados
cm <- table(results,data_test$results)
#View(cm)

# Procentaje de aciertos (Precision)
overall_accuracy <- sum(diag(cm)) / sum(cm)
p_compensated <- sum(cm[1,1]) / sum(cm[1,])
p_negative <- sum(cm[2,2]) / sum(cm[2,])
p_primary <- sum(cm[3,3]) / sum(cm[3,])
```