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

INFORME EXPERIENCIA 4 ANÁLISIS DE DATOS

CLASIFICADOR BAYESIANO

Integrantes:

Marcela Rivera Castro

Kevin Alvarez Alvial

Profesor:

Max Chacón

Ayudante:

Adolfo Guzmán

Santiago - Chile

18 de abril de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

Antiguamente cerca de los años 1880, se tenía completo desconocimiento a cerca de la glándula tiroides y sus funciones. No existía advertencia de lo importante que es para el organismo humano. Entre los conocimientos que se manejaban en aquel momento, se sabía del cretinismo y los casos de mixedema del adulto de Gull, sin embargo no se conocía su origen tiroideo.

En 1883 el cirujano Teodoro Emilio Kocher, realizó una publicación sobre las consecuencias funestas de la tiroidectomía radical. Resolviendo en 1888 que cretinismo, mixedema y "caquexia" postiroidectomía eran síndromes estrechamente relacionados, si no idénticos, y se debían los tres a la pérdida de la función tiroidea.(Aguirre, 2002)

En la actualidad se conocen muchas de las enfermedades relacionadas con la tiroides. En el presente informe se abordará el hipotiroidismo. Hipotiroidismo significa "poca hormona tiroidea". Ocurre cuando la glándula tiroidea esta dañada y no es capaz de producir las hormonas tiroideas suficientes para mantener el metabolismo del cuerpo normal. El exceso de TSH puede causar que la glándula tiroidea aumente de tamaño lo que se llama bocio. Existen otras causas de hipotiroidismo como las tiroiditis autoinmunes o virales que pueden generar el mismo cuadro final pero sin bocio.(de Endocrinología Facultad de Medicina UC, s.f.)

1.2 ORGANIZACIÓN DEL DOCUMENTO

El documento consta de cuatro secciones principales: marco teórico para entender el dominio del problema, obtención de reglas seguido de su respectivo análisis de resultados y finalmente conclusiones, donde se indica el aprendizaje obtenido a partir del desarrollo de la experiencia.

1.3 METODOLOGÍAS Y HERRAMIENTAS UTILIZADAS

- Para el estudio de los datos se utilizará el programa R studio.
- La base de datos a utilizar es: allhypo.data, allhypo.test y allhypo.names

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.

El algoritmo de clasificación bayesiano ingenuo de ISW es un clasificador probabilístico. Se basa en modelos de probabilidad que incorporan fuertes supuestos de independencia. (Center, s.f.)

2.2 PRIORIDAD A PRIORI

La probabilidad a priori p(ci) será la probabilidad de que el sujeto se clasifique en la clase c_i. Se representa como:

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

2.3 PRIORIDAD A POSTERIORI

Especifica la probabilidad de que el sujeto pertenezca a la clase C_i dado un valor x. Tener el valor de x dependerá del hecho posterior de que la variable de características x sea medida (Chacón, 2017). Para dos clases se tiene:

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

Para obtener el valor de p(c_i/x), se requiere conocer las relaciones de probabilidades condicionales.

CAPÍTULO 3. OBTENCIÓN DEL CLASIFICADOR

3.1 PRE-PROCESAMIENTO

Dado a que en el dataset utilizado para el estudio existen variedad de datos tomados de forma aleatoria, es necesario realizar un análisis previo para verificar que los datos sean consistentes y no hayan anomalías que afecten al resultado del estudio.

3.1.1 Eliminación de registros NA

Dentro de los datos del dataset puede ocurrir que algunos de éstos datos sean nulos para ciertas variables, es decir, que no fueron tomados o que posteriormente fueron eliminados, éstos datos son representados en el dataset como "?" y es a lo que se llama registro NA o desconocido. En este caso, si una observación tiene un registro NA, ésta se elimina.

3.1.2 Eliminación de variables

La eliminación de variables puede darse debido a razones variadas, pero específicamente, para este caso el criterio utilizado, es que la variable no contenga ningún dato, que contenga solo registros desconocidos (NA) o que para el objetivo del estudio ésta no entregue mucha información.

3.1.2.1 TBG

Esta variable dentro del dataset tiene mediciones que no fueron realizadas, ya que todos los datos son registros desconocidos, dado esto, la variable no entrega información y por ende se puede eliminar.

3.1.2.2 Variables de medición

En el dataset, existen variables booleanas para señalar si un examen de hormona fue realizado o no, esta variable es usada para filtrar los registros NA, dado a que si es falsa, significa que el dato es desconocido, éstas variables son:

TSH measured, **T3 measured**, **TT4 measured**, **T4U measured**, **FTI measured**.

Estas variables por si mismas y luego del filtro, no entregan información útil para el objetivo del estudio, por lo que se tornan innecesarias y pueden ser eliminadas.

3.1.2.3 Fuente de referencia

Esta variable indica la fuente de la cual se obtuvo los datos de un sujeto (u observación) en particular, por lo que esta variable no entrega información útil, por lo que se puede

eliminar.

3.2 CLASE

Una vez realizado el pre-procesamiento de datos, es necesario proceder con decidir cual debe ser la clase a utilizar. Una forma sencilla de decidir esto, es plantear que es lo que se busca. En este caso, es importante conocer cuales son las personas que padecen de hipotiroidismo, para esto la pregunta sería: ¿Cuál es la probabilidad de que una persona padezca de hipotiroidismo a partir de los resultados de sus exámenes hormonales, edad, sexo, entre otros?.

A partir de la pregunta planteada, se ha llegado a la conclusión de que la variable que mejor lo describe es clasification, la cual se denomina como atributo predictor, ya que ésta es capaz de definir con mejor aproximación que tan relacionada puede estar la enfermedad con respecto a algunos resultados en particular. Esto se ha decidido así, dado que después de una larga investigación con respecto a los síntomas de esta enfermedad, se deduce que no hay un único atributo capaz de definir si un paciente padece o no de hipotiroidismo, ya que los atributos capaces de condicionar esta enfermedad son varios, dentro de los cuales se encuentran las hormonas como T3, TT4, entre otros.

Cabe destacar que el clasificador Bayesiano ingenuo se basa en que los atributos son totalmente independientes entre sí, por lo cual este método se basa a partir de información previa o de evidencia (Richter-Walsh, 2017). A partir de esto, se ve la necesidad de obtener una base de datos de entrenamiento la cual debe ser representativa con respecto al tema en estudio, por lo cual se ha decidido utilizar la base de datos allhypo.data. Por otro lado, se requiere de una base de datos de prueba, por lo que en este estudio se utilizará allhypo.test ya que esta pertenece al mismo estudio de hipotiroidismo, así que se asume que los datos son totalmente representativos ya que en esta base de datos se encuentra un conjunto de datos que contemplan los mismos atributos, clasificando a los individuos estudiados según su condición, sin embargo hacer uso de esta base de datos permite obtener mejores conclusiones acerca de la eficiencia del clasificador bayesiano ingenuo ya que se tiene un mejor punto de comparación.

Una vez definido el atributo predictor y la base de datos de entrenamiento, es posi-

ble hacer uso del clasificador bayesiano, pudiendo estudiar la precisión de éste para el conjunto de datos en estudio.

3.3 CONJUNTOS UTILIZADOS

Finalmente, la cantidad de atributos a utilizar son 22, los cuales son:

1. age (variable continua)
2. sex (variable categórica)
3. on thyroxine (variable categórica)
4. query on thyroxine (variable categórica)
5. on antithyroid medication (variable categórica)
6. sick (variable categórica)
7. pregnant (variable categórica)
8. thyroid surgery (variable categórica)
9. I131 treatment (variable categórica)
10. query hypothyroid (variable categórica)
11. query hyperthyroid (variable categórica)
12. lithium (variable categórica)
13. goitre (variable categórica)
14. tumor (variable categórica)
15. hypopituitary (variable categórica)
16. psych (variable categórica)
17. TSH (variable continua)

18. T3 (variable continua)

19. TT4 (variable continua)

20. T4U (variable continua)

21. FTI (variable continua)

22. clasification (ATRIBUTO PREDICTOR)

Utilizando estos atributos, se tiene que el total de datos para la base de datos de entrenamiento es de 1947.

Por otro lado, el conjunto de prueba contiene las 22 variables mencionadas anteriormente, sin embargo, se tienen 696 observaciones, por lo que es un conjunto con menor cantidad de datos pero que de igual manera permitirá obtener un mejor análisis de la eficiencia de éste clasificador.

CAPÍTULO 4. ANÁLISIS DE RESULTADOS

4.1 CLASIFICACIÓN

Luego de realizado el proceso de clasificación con los datos de prueba a través del clasificador entrenado, se pueden observar los resultados a través de una matriz de confusión, donde se compara la predicción de cada clase contra las instancias de clases reales, o del set ingresado, en este caso, el conjunto de datos de prueba.

En la tabla 4.1 se puede observar la matriz de confusión del clasificador bayesiano ingenuo con los resultados de la clasificación, las instancias de clases indicadas en las filas corresponden a la clasificación realizada, y en las columnas corresponde a la clase real, perteneciente a los datos de prueba.

Tabla 4.1: Matriz de confusión del clasificador.

Resultados	**compensated hypothyroid**	**negative**	**primary hypothyroid**
compensated hypothyroid	11	1	1
negative	19	634	3
primary hypothyroid	0	2	25
secondary hypothyroid	0	0	0

Un detalle que puede notarse es que la instancia de clase ”secondary hypothyroid”solo se encuentra contemplada en la clase del clasificador y no en la clase real, esto es debido a que esta instancia es considerada en el modelo dado ya que el conjunto de entrenamiento contiene observaciones con esta clasificación, pero no así el conjunto de entrenamiento, el cual no presentaba observaciones clasificadas de esta forma, como resultado la clasificación no presentó casos que fuesen contemplados dentro de la instancia, pero no se descarta que de otros conjuntos, algunas observaciones puedan ser clasificadas de esta forma aunque no la contenga el conjunto a clasificar, incurriendo en un error por parte del clasificador.

De la matriz de confusión se puede desprender medidas de rendimiento del clasificador en relación a su precisión, entre éstas se pueden encontrar índices para medir la precisión general, y valores predictivos para cada clase, en la tabla 4.2 se encuentran éstos índices

calculados a partir de los resultados mostrados en la matriz de confusión.

Tabla 4.2: Índices de precisión del clasificador.

Medida	Proporción	Porcentaje
Precisión general	0.9626437	96.26 %
Valor predictivo: negative	0.9664634	96.65 %
Valor predictivo: primary hypothyroid	0.9259259	92.59 %
Valor predictivo: compensated hypothyroid	0.8461538	84.62 %

Estas medidas muestran que el clasificador tiene una alta precisión, ya que en relación a la clase real, clasificó correctamente, en general, un 96.26 %, es decir, solo cometió un error de clasificación en el 3.74 % de las observaciones, así mismo, en relación a las instancias de clase, negative, primary hypothyroid y compensated hypothyroid, cometió un error de 3.35 %, 7.41 % y 15.38 % respectivamente.

4.2 COMPARACIÓN CON EXPERIENCIAS ANTERIORES

Es posible realizar una comparación entre los métodos de experiencias anteriores y el utilizado en esta instancia, con el objetivo de observar similitudes y diferencias en éstos.

4.2.1 k-medias

No existe realmente punto de comparación entre el método de clustering de k-medias y el clasificador bayesiano ingenuo dado a que los métodos difieren completamente y tienen objetivos diferentes, k-medias es un método de aprendizaje no supervisado que tiene como objetivo agrupar las observaciones basándose en la distancia entre ellas, mientras que el clasificador bayesiano ingenuo, siendo un método de aprendizaje supervisado, hace énfasis en las probabilidades a priori y posteriori (teorema de Bayes) y bajo esto hace la clasificación con la instancia de clase que tiene máxima probabilidad para una observación, que sea ingenuo significa que se basa además en un supuesto de independencia entre las variables.

En relación a los resultados obtenidos, con k-medias en primera instancia se logró agrupar por sexo a los sujetos, lo cual no era el resultado esperado, luego de filtrar variables, solo se dejaron las que, según la teoría, eran más impactantes para determinar el

cuadro de hipotiroidismo, como resultado se logró agrupar de acuerdo a las medidas de TSH y T3, TT4, TU4 y FTI una tendencia del paciente a tener hipotiroidismo. Por parte de la experiencia actual, al ser supervisado el método, se enfocó la clasificación de acuerdo al cuadro del paciente, logrando resultados exitosos en relación a la clasificación misma y la precisión del clasificador.

4.2.2 Reglas de asociación

Para este caso, ambos métodos son supervisado y ambos son enfocados en la variable predicha o clase, con el objetivo de predecir el cuadro de un sujeto de acuerdo a las variables involucradas, aún así, difieren en su metodología. Reglas de asociación busca la relación entre las variables de tal forma que la asociación de un conjunto de ellas tiene como consecuencia la clase, entonces en este método es posible observar las diferentes variables que son asociadas, ver su importancia en el modelo y la clasificación que realizan para una observación, a diferencia de esto, el clasificador bayesiano ingenuo, basado en las probabilidades, optimiza la probabilidad a posteriori pero siendo un proceso de caja negra, por lo cual no es posible determinar que variables tienen mayor importancia en el modelo, solo es posible determinar la clase de las observaciones.

Dado a que los objetivos de ambos métodos difieren, los resultados igualmente lo hacen, con el método de reglas de asociación se hallaron las reglas que tienen con consecuente hipotiroidismo, pero sin especificar el tipo, esto se hizo con el objetivo de hallar las variables mas importantes y que entregan mayor información al modelo, con el clasificador bayesiano ingenuo, se busca simplemente clasificar a los sujetos de acuerdo a las instancias establecidas, por lo que los resultados obtenidos son en base a la clasificación realizada y la precisión del clasificador para realizar esta tarea.

CAPÍTULO 5. CONCLUSIONES

En la experiencia se realizó un estudio sobre clasificación de sujetos de acuerdo al cuadro de hipotiroidismo, para ello se utilizó un método de aprendizaje supervisado basado en probabilidades del teorema de Bayes, llamado clasificado Bayesiano ingenuo, el cual busca optimizar la probabilidad a posteriori de las variables.

Para lograr el objetivo y clasificar las observaciones, primero se realizó un preprocesamiento de los datos con tal de eliminar todos los que son innecesarios o que pudieran hacer ruido, y de esta forma lograr mejores resultados.

En relación a los resultados obtenidos, el clasificador logró una precisión de 96.26 % en la clasificación de las observaciones de conjunto de prueba, por lo cual se puede decir que el sistema aprendió correctamente y logró realizar un modelo representativo a partir del conjunto de entrenamiento, por lo cual este clasificador podría ser usado para clasificar nuevos datos que sean tomados de nuevos sujetos, siempre y cuando las variables medidas sean concordantes con las que conoce el clasificador.

En comparación a las experiencias anteriores, la clasificación bayesiana es mucho más precisa y práctica que los métodos anteriormente estudiados, en comparación con el algoritmo de clustering k-medias, éste al ser supervisado se puede enfocar en la clasificación con relación a una variable específica, la cual será predicha, por lo que es más práctico para problemas más específicos y más complejos en dimensionalidad. En comparación con las reglas de asociación, ambos siendo apredizajes supervisados, son enfocados en la predicción de una variable específica, la diferencia principal entre estos métodos reside en su metodología, mientras que el clasificador bayesiano se basa en probabilidades, maximizando la probabilidad a posteriori, el método de reglas de asociación busca asociación entre variables que tengan como consecuencia la variable a predecir, entonces, un método al basarse en estadística no muestra el procedimiento de clasificación, mientras que el otro presenta las reglas, donde es posible observar las variable involucradas o que tienen mayor importancia para el modelo, lo cual es información valiosa para entender como se realizó el proceso de clasificación.

A modo de aprendizaje, se conoció un nuevo método de aprendizaje supervisado, ba-

sado en estadística, mas específicamente en el teorema de Bayes y probabilidades a priori y posteriori, donde la clasificación se realiza maximizando esta última, como no se obtiene información sobre el procedimiento para concluir la clase de una observación, se podría decir que éste es un método de caja negra, lo cual lo diferencia del clasificador anteriormente estudiado que usa el método de reglas de asociación, donde se pueden observar las reglas, variables importantes y medidas de confianza de las reglas, sin embargo con una clasificación no tan exhaustiva, ya que solo se utilizó una clasificación binaria, gracias a esto se pudo observar las principales diferencias entre el aprendizaje por método de caja negra y el aprendizaje orientado al conocimiento.

Cabe destacar que el clasificador bayesiano resultó muy útil, ya que a pesar de basarse en un principio que determina que las variables son totalmente independientes entre sí, se puede de igual manera obtener resultados acorde a lo esperado independiente del problema, ya que lo fundamental es saber escoger el atributo predictor para así obtener resultados coherentes y robustos. Además su implementación es sencilla ya que permite al programador abstraerse de muchos procesamientos de datos que pueden resultar engorrosos, como calcular todas las probabilidades apriori y posteriori.

CAPÍTULO 6. BIBLIOGRAFÍA

Aguirre, C. P. (2002). Emil Theodor Kocher (1841-1917). Recuperado desde http://www.historiadelamedicina.org/kocher.html

Center, I. K. (s.f.). Bayesiano ingenuo ISW. Recuperado desde https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.0.0/modeler_mainhelp_client_ddita/clementine/dbmining_ibm_naivebayes.html

Chacón, M. (2017). Clasificación Bayesiana. Recuperado desde http://www.udesantiagovirtual.cl/moodle2/pluginfile.php?file=%5C%2F217599%5C%2Fmod_resource%5C%2Fcontent%5C%2F1%5C%2FCap%5C%C3%5C%ADtulo%5C%20VI%5C%20An%5C%C3%5C%A1lisis%5C%20de%5C%20Datos_CB_N.pdf

de Endocrinología Facultad de Medicina UC, D. (s.f.). HIPOTIROIDISMO. Recuperado desde http://redsalud.uc.cl/ucchristus/VidaSaludable/Glosario/H/hipotiroidismo.act

Richter-Walsh, S. (2017). Clasificación Naive Bayes en R (Parte 2). Recuperado desde https://www.r-bloggers.com/naive-bayes-classification-in-r-part-2/

CAPÍTULO 7. ANEXO: CÓDIGO EN R

```
library("e1071")


preprocessing <- function(rawdata){
  #delete id from clasification|id column
  d <- c()
  for(i in rawdata$`clasification|id`){
    d <- c(d, strsplit(i, ".", fixed = TRUE)[[1]][1])
  }
  colnames(rawdata)[30] <- "clasification"
  rawdata$clasification <- as.factor(d)
  #data pre-processing
  #delete NA values
  data <- rawdata[(rawdata$age!="?" & rawdata$sex!="?" & rawdata$on.
      thyroxine!="?" & rawdata$query.on.thyroxine!="?" & rawdata$on.
      antithyroid.medication!="?" & rawdata$sick!="?" & rawdata$pregnant
      !="?" & rawdata$thyroid.surgery!="?" & rawdata$I131.treatment!="?"
       & rawdata$query.hypothyroid!="?" & rawdata$query.hyperthyroid!="?
      " & rawdata$lithium!="?" & rawdata$goitre!="?" & rawdata$tumor!="?
      " & rawdata$hypopituitary!="?" & rawdata$psych!="?" & rawdata$TSH.
      measured!="f" & rawdata$T3.measured!="f" & rawdata$TT4.measured!="
      f" & rawdata$T4U.measured!="f" & rawdata$FTI.measured!="f"),]
  #delete variable TBG
  data$TBG.measured <- NULL
  data$TBG <- NULL
  #delete measuring variables
  data$TSH.measured <- NULL
  data$T3.measured <- NULL
  data$TT4.measured <- NULL
  data$T4U.measured <- NULL
  data$FTI.measured <- NULL
  #delete referral.source variable
  data$referral.source <- NULL
```

```
  #data format transform
  #nominal variables
  data$sex <- as.factor(data$sex)
  data$on.thyroxine <- as.factor(data$on.thyroxine)
  data$query.on.thyroxine <- as.factor(data$query.on.thyroxine)
  data$on.antithyroid.medication <- as.factor(data$on.antithyroid.
      medication)
  data$sick <- as.factor(data$sick)
  data$pregnant <- as.factor(data$pregnant)
  data$thyroid.surgery <- as.factor(data$thyroid.surgery)
  data$I131.treatment <- as.factor(data$thyroid.surgery)
  data$query.hypothyroid <- as.factor(data$query.hypothyroid)
  data$query.hyperthyroid <- as.factor(data$query.hyperthyroid)
  data$lithium <- as.factor(data$lithium)
  data$goitre <- as.factor(data$goitre)
  data$tumor <- as.factor(data$tumor)
  data$hypopituitary <- as.factor(data$hypopituitary)
  data$psych <- as.factor(data$psych)

  #continuous variables
  data$age <-as.numeric(data$age)
  data$TSH <- as.numeric(data$TSH)
  data$T3 <- as.numeric(data$T3)
  data$TT4 <- as.numeric(data$TT4)
  data$T4U <- as.numeric(data$T4U)
  data$FTI <- as.numeric(data$FTI)

  return(data)
}

#Data read
rawdata_train <- read.csv("allhypo.data", header = FALSE, sep = ",",
    stringsAsFactors = FALSE)
```

```
colnames(rawdata_train) <- 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", "clasification|id")

rawdata_test <- read.csv("allhypo.test", header = FALSE, sep = ",",
    stringsAsFactors = FALSE)
colnames(rawdata_test) <- 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", "clasification|id")

#Preprocessing for both data, train and test
data_train <- preprocessing(rawdata_train)
data_test <- preprocessing(rawdata_test)

#Training
model <- naiveBayes(clasification ~., data = data_train)

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

#Confusion Matrix - Predicted vs Trained
cm <- table(results,data_test$clasification)

#Accuracy measures
overall_accuracy <- sum(diag(cm)) / sum(cm)
pv_compensated <- sum(cm[1,1]) / sum(cm[1,])
```

```
pv_negative <- sum(cm[2,2]) / sum(cm[2,])
pv_primary <- sum(cm[3,3]) / sum(cm[3,])
```