Aprendizaje Automático Taller 1

Andres Racines Bolaños.
Mestria En Ingenieria
Pontifiacia Universidad Javeriana
Cali-Colombia
Email:racines@javerianacali.edu.co

***Resumen*- En este informe se documentan los resultados del Actividades realizadas en el primer taller del curso aprendizaje automático: "Pre-procesamiento y técnicas de aprendizaje simbólico".**

I. Explicar los siguientes conceptos

- Maldición de la dimensionaldad

  Es un problema habitual que se presenta Minería de datos y consiste en que al tratar de tener todas las características o atributos posibles para describir un sistema, lo que podría suponerse como la mejor estrategia para realizar una correcta clasificación o un eficiente trabajo de Minería de datos, sin embargo esto causara altos costos de procesamiento además de selección de datos irrelevantes o redundantes la maldición de la dimensionalidad consiste en que a mayor número de atributos que describan el problema el tamaño de conjunto de datos también crecerá de manera exponencial.

- Sobre entrenamiento

  El sobre entrenamiento ocurre cuando un clasificador se ajusta muy bien al conjunto de entrenamiento pero falla al intentar clasificar un conjunto nuevo de datos.

- Sesgo de aprendizaje

  Un margen de predicción significativamente distinto de cero revela que hay un error en alguna parte del modelo, ya que indica que el modelo es incorrecto con respecto a la frecuencia con la que ocurren las etiquetas positivas, cuando se promedian las probabilidades predichas de cada clase, el resultado debería ser 1%. La predicción promedio del modelo es del 20% de probabilidad, podemos concluir que muestra sesgo de predicción.

  Las posibles causas fundamentales del sesgo de predicción son las siguientes:

  1. Conjunto de atributos incompleto
  2. Conjunto de datos contaminado
  3. Canalización con errores
  4. Muestra de entrenamiento con sesgo
  5. Regularización excesivamente fuerte

II. Potencial de Aprendizaje automático

El aprendizaje automático tiene grandes potenciales a futuro aunque hoy es una realidad que trasciende diferentes campos el maketing y la medicina, en la actualidad se ha ocupado de reducir el tiempo en el que se crean pronósticos precisos y fiables y ha tenido impacto en la capacidad de las empresas para planificar, hacer presupuestos, identificar anomalías en grandes conjuntos de datos, etc. Cada vez es más frecuente la incursión del aprendizaje automático diferentes áreas, remplazando algunas actividades que realizadas por humanos hasta el momento esto supone un gran riesgo para la dinámica de la economía actual, algunas profesiones comenzarían a desaparecer como los abogado y los médicos las cuales tienen como principal tarea el diagnostico o clasificación de acuerdo a unas evidencias, en la tecnología podría ocurrir que un virus informático creado por un ser humano sea aprendido por máquina y para generar nuevos virus con mayor potencial.

III. Pre-procesamiento

Las actividades de pre-procesamiento fueron realizadas con el conjunto de datos allrep el cual es un registro Registros de enfermedades tiroideas suministrados por el Instituto Garavan y J. Ross ; Quinlan, Instituto de Nueva Gales del Sur, Syndney, Australia del años 1987, se realizaron las actividades de pre-procesamiento a dos conjunto de datos allrep-data y allrep-test.

Se descargo el conjunto de datos de entrenamiento allrep-data el cual contiene 30 atributos y 2800 registro.

```
> dim(train)
[1] 2800   30
>
```

Con el comando head(train) se verifican las columnas del conjunto de datos y se identifica que la columna 28 esta totalmente vacía , se procede a eliminar la columna.

```
V23  V24 V25 V26 V27 V28   V29                        V30
  t 1.14   t 109   f   ?  SVHC              negative.|3733
  f    ?   f   ?   f   ? other              negative.|1442
  t 0.91   t 120   f   ? other              negative.|2965
  f    ?   f   ?   f   ? other replacement therapy.|806
  t 0.87   t  70   f   ?   SVI              negative.|2807
  t  1.3   t 141   f   ? other              negative.|3434
```

```
#SE ELIMINA LA COLUMNA 28 QUE ESTA TOTALMENTE VACIA
train<-data.frame(train[-28])
> dim(train)
[1] 2800   29
>
```

Se elimienan los valores nulos

```
delete.dirt <- function(DF, dart=c('?')) {
  dirty_rows <- apply(DF,1, function(r) !any(r %in% dart))
  DF <- DF[dirty_rows, ]
}
train<- delete.dirt(train)
```

Una vez eliminado los registro nulos el dataset queda con el siguiente tamaño.

```
> dim(train)
[1] 1947   29
>
```

Se verifica el tipo de datos que contiene cada uno de los atributos de dataset

```
> sapply(train, class)
       V1        V2        V3        V4        V5        V6        V7        V8
 "factor"  "factor"  "factor"  "factor"  "factor"  "factor"  "factor"  "factor"
       V9       V10       V11       V12       V13       V14       V15       V16
 "factor"  "factor"  "factor"  "factor"  "factor"  "factor"  "factor"  "factor"
      V17       V18       V19       V20       V21       V22       V23       V24
 "factor"  "factor"  "factor"  "factor"  "factor"  "factor"  "factor"  "factor"
      V25       V26       V27       V29       V30
 "factor"  "factor"  "factor"  "factor"  "factor"
>
```

Como es posible observar todos los atributos del dataset son del tipo factor, se procede a convertirlos en numéricos, se convierten a numéricos excepto la columna que contiene la clase V30

```
> head(train)
   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17  V18 V19 V20
1  41  2  1  1  1  1  1  1  1   1   1   1   1   1   1   1   2  1.3   2 2.5
5  70  2  1  1  1  1  1  1  1   1   1   1   1   1   1   1   2 0.72   2 1.2
8  80  2  1  1  1  1  1  1  1   1   1   1   1   1   1   1   2  2.2   2 0.6
9  66  2  1  1  1  1  1  1  1   1   1   1   1   2   1   1   2  0.6   2 2.2
10 68  3  1  1  1  1  1  1  1   1   1   1   1   1   1   1   2  2.4   2 1.6
11 84  2  1  1  1  1  1  1  1   1   1   1   1   2   1   1   2  1.1   2 2.2
   V21 V22 V23  V24 V25 V26 V27 V29          V30
1    2 125   2 1.14   2 109   1   3 negative.|3733
5    2  61   2 0.87   2  70   1   5 negative.|2807
8    2  80   2  0.7   2 115   1   5 negative.|1367
9    2 123   2 0.93   2 132   1   5 negative.|1787
10   2  83   2 0.89   2  93   1   5 negative.|2534
11   2 115   2 0.95   2 121   1   5 negative.|1485
```

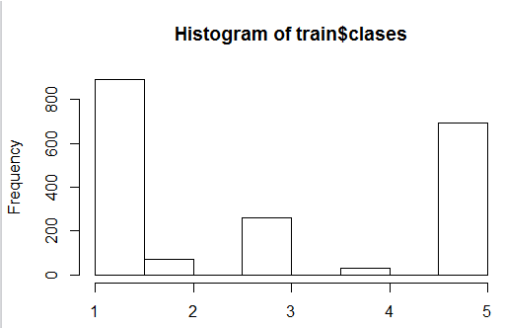
Se agregan los etiquetas a los direntes atributos

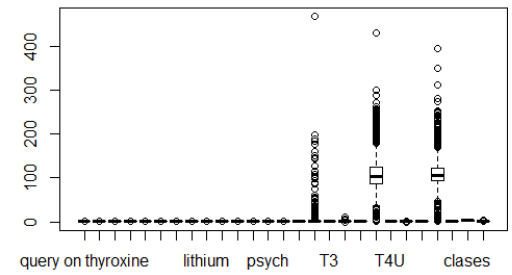
```
names(train)<-c("age","sex","on thyroxine","query on thyroxine","on antithyroid medication",
"sick","pregnant","thyroid surgery","I131 treatment","query hypothyroid","query hyperthyroid",
"lithium","goitre","tumor","hypopituitary","psych","TSH measured","TSH","T3 measured",
"T3","TT4 measured","TT4","T4U measured","T4U","FTI measured","FTI","TBG measured",
"referral source","clases")
```

media, mediana y moda, además de los valores mínimos y máximos por cada atributo

```
> summary(train)
      age               sex          on thyroxine   query on thyroxin
 Min.   :   1.0   Min.   :  1.00   Min.   :2.000   Min.   :1.000
 1st Qu.: 689.5   1st Qu.: 37.00   1st Qu.:2.000   1st Qu.:1.000
 Median :1377.0   Median : 55.00   Median :2.000   Median :1.000
 Mean   :1387.3   Mean   : 53.22   Mean   :2.342   Mean   :1.094
 3rd Qu.:2085.5   3rd Qu.: 69.00   3rd Qu.:3.000   3rd Qu.:1.000
 Max.   :2800.0   Max.   :455.00   Max.   :3.000   Max.   :2.000
 thyroid surgery I131 treatment  query hypothyroid query hyperthyroi
 Min.   :1.000   Min.   :1.000   Min.   :1.000     Min.   :1.000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000     1st Qu.:1.000
 Median :1.000   Median :1.000   Median :1.000     Median :1.000
 Mean   :1.016   Mean   :1.013   Mean   :1.016     Mean   :1.049
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000     3rd Qu.:1.000
```

Se grafica un histograma con la columna que contiene la clases para verificar que las barras de histograma correspondan a las 4 clases, que describen el problema. “replacement therapy, underreplacement, overreplacement, negative.”





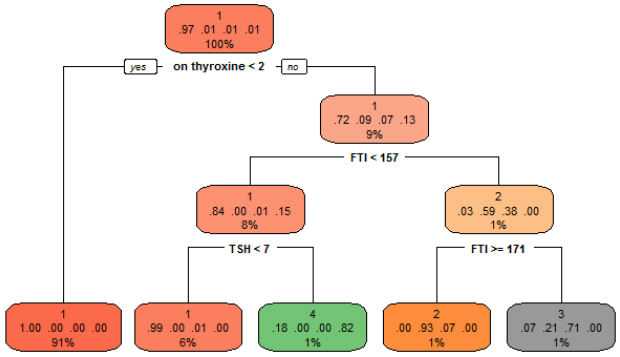
Se realiza el mismo procedimiento con el conjunto test

IV. Técnicas de aprendizaje simbólico

La técnica de aprendizaje simbólico escogida para realizar el taller son los arboles de decisión, primero se crea un modelo de arbol con los datos de entrenamiento

```
arbol_1 <- rpart(clases~., data = train, method = 'class')
```

Posteriormente se utiliza rpart.plot para graficar. El árbol creado en el punto anterior



Posterior a se procede a generar la matrix de confusión en la cual se puede observar el rendimie

```
> confusionMatrix(predict_1,test$clases)
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 666   0   1   0
         2   0   7   3   0
         3   3   1   1   0
         4   4   0   0  10

Overall Statistics

               Accuracy : 0.9828
                 95% CI : (0.9701, 0.9911)
    No Information Rate : 0.967
    P-Value [Acc > NIR] : 0.008286

                  Kappa : 0.7633
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.9896  0.87500 0.200000  1.00000
Specificity            0.9565  0.99564 0.994211  0.99417
Pos Pred Value         0.9985  0.70000 0.200000  0.71429
Neg Pred Value         0.7586  0.99854 0.994211  1.00000
Prevalence             0.9670  0.01149 0.007184  0.01437
Detection Rate         0.9569  0.01006 0.001437  0.01437
Detection Prevalence   0.9583  0.01437 0.007184  0.02011
Balanced Accuracy      0.9731  0.93532 0.597106  0.99708
>
```