SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms

Introducción

En esta PEC vamos a realizar un informe basado en los datos del artículo:

D Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318

Los datos se pueden obtener del dataset "Cardiotocography" desde la *UCI Machine Learning Repository*, disponible en http://archive.ics.uci.edu/ml. Aunque también se ha dejado en la PEC el documento "CTG.csv" con los datos.

En el trabajo se procesaron automáticamente 2126 cardiotocogramas fetales (CTG) y se midieron 21 variables para caracterizar el diagnostico de "foetal heart rate" (FHR). Los CTG fueron clasificados por tres especialistas expertos, y se realizo la clasificación por consenso para cada uno de los CTG. La clasificación fue tanto con respecto a un patrón morfológico (A, B, C. ...) como a un estado fetal (N, S, P).

Información de cada variable es:

LB Línea de base de la FHR (latidos por minuto)

AC Número de aceleraciones por segundo

FM Nº de movimientos fetales por segundo

UC Nº de contracciones uterinas por segundo

DL Nº de deceleraciones de luz por segundo

DS Nº de deceleraciones severas por segundo

DP Nº de deceleraciones prolongadas por segundo

ASTV porcentaje de tiempo con variabilidad anormal a corto plazo

MSTV valor medio de la variabilidad a corto plazo

ALTV porcentaje de tiempo con variabilidad anormal a largo plazo

MLTV valor medio de la variabilidad a largo plazo

Width anchura del histograma de FHR

Min mínimo de histograma de FHR

Max máximo de histograma de FHR

Nmax Nº de picos de histograma

Nzeros Nº de ceros del histograma

Mode moda del histograma

Mean media del histograma

Median mediana del histograma

Variance varianza del histograma

Tendency tendencia del histograma (-1 = asimétrico a la izquierda, 0 = simétrico, 1 = asimétrico a la derecha)

CLASS Código de clase de patrón FHR (1 a 10, que corresponde a las clases A, B, C, D, E, AD, DE, LD, FS y SUSP, respectivamente)

NSP código de clase del estado fetal (N = normal, S = sospechoso, P = patológico)

Significado de cada código de clase de patrón FHR:

- **A** calm sleep
- **B** REM sleep
- **C** calm vigilance
- **D** active vigilance
- **SH** shift pattern (A or Susp with shifts)
- **AD** accelerative/decelerative pattern (stress situation)
- **DE** decelerative pattern (vagal stimulation)
- **LD** largely decelerative pattern
- **FS** flat-sinusoidal pattern (pathological state)
- **SUSP** suspect pattern

Objetivo:

En esta PEC se usan los datos del artículo para **implementar** los diferentes **algoritmos estudiados**: *k-Nearest Neighbour, Naive Bayes, Artificial Neural Network, Support Vector Machine, Arbol de Decisión y Random Forest* para **predecir** los tres tipos de **diagnóstico del estado fetal**: normal, sospechoso y patológico.

Puntos importantes:

1. Observar que la variable CLASS esta fuertemente asociada a la variable NSP. Ambas no son variables obtenidas del CTG si no deducidas por los especialistas, por tanto la variable CLASS no puede ser una variable explicativa. Así que solo hay que considerar como variables explicativas las 21 primeras variables.

2. En cada algoritmo hay que realizar las siguientes tres etapas: 1) Transformacion de los datos (en caso necesario) 2) Entrenar el modelo 3) Predicción y Evaluación del algoritmo. En la fase 3) probar diferentes parámetros del algoritmo para posteriormente evaluar su rendimiento.

3. Se debe aplicar la misma selección de datos training y test en todos los algoritmos. Utilizando la semilla aleatoria `12345`, para separar los datos en dos partes, una parte para training (67%) y otra parte para test (33%). Si se prefiere, se puede escoger otro tipo de partición de los datos para hacer la selección de training y test como por ejemplo k-fold crossvalidation, boostrap, random splitting, etc. Lo que es importante es mantener la misma selección para todos los algoritmos.

4. En todos los casos se evalua la calidad del algoritmo con la información obtenida de la función `confusionMatrix()` del paquete `caret`.

5. Para la ejecución especifica de cada algoritmo se puede usar la función de cada algoritmo como se presenta en el libro de referencia o usar el paquete `caret` con los diferentes modelos de los algoritmos. O incluso, hacer una versión mixta.

6. Comentario sobre el informe dinámico. Una opción interesante del knitr es poner `cache=TRUE`. Por ejemplo:

   ```
   knitr::opts_chunk$set(echo = FALSE, comment = NULL, cache = TRUE)
   ```

   Con esta opción al ejecutar el informe dinámico crea unas carpetas donde se guardan los resultados de los procesos. Cuando se vuelve a ejecutar de nuevo el informe dinámico solo ejecuta codigo R donde se ha producido cambios, en el resto lee la información previamente descargada. Es una opción muy adecuada cuando la ejecución es muy costosa computacionalmente.

Informe de la PEC

Las soluciones se presentarán mediante un informe dinámico R markdown con la siguiente estructura:

1. Titulo: igual que el de la PEC, autor, fecha de creación e indice de apartados de la PEC.
2. Sección de lectura, exploración de los datos y obtención de los muestras de train y test. Recordar que un primer paso es, si hace falta, transformar las variables leidas al tipo de objeto R adecuado al tipo de variable. La exploración de los datos se aplica a todas las variables leidas. (*Puntuación*: 10%)
3. Sección de aplicación de cada algoritmo para la clasificación. Está formado por subsecciones que corresponden a cada algoritmo: k-Nearest Neighbour, Naive Bayes, Artificial Neural Network, Support Vector Machine, Arbol de Decisión y Random Forest manteniendo este orden. (*Puntuación*: 60%)

   En cada algoritmo hay que realizar las tres etapas mencionadas anteriormente.
4. Sección de conclusión y discusión sobre el rendimiento, interpretabilidad, ... de los algoritmos para el problema tratado. Proponer que modelo o modelos son los mejores. (*Puntuación*: 20%)

Un caracteristica que se valorará es hasta que punto es el informe "dinámico". En el sentido de adaptarse el informe a cambios en los datos. (*Puntuación*: 10%)

Se entregan tres ficheros:

1. Fichero ejecutable (.Rmd) que incluya un texto explicativo que detalle los pasos implementados en el script y el código de los análisis. NOTA: Para facilitar la ejecución, no usar un ruta fija para la lectura del fichero, asociarlo al area de trabajo donde este el fichero .Rmd.
2. Informe (pdf) resultado de la ejecución del fichero Rmd anterior.
3. El fichero de datos CTG.csv.