Proyecto Final. Cardiotocography

Anabel Gómez Ríos y Gustavo Rivas Gervilla

16 de junio de 2016

1. Definición del problema a resolver y enfoque elegido.

En este proyecto vamos a trabajar con una base da datos algo mayor que las que hemos venido usando en las prácticas (2126 instancias con 23 atributos cada una) con el objetivo de poner en práctica los conocimientos adquiridos en la asignatura para resolver un problema de clasificación del mundo real.

La base de datos elegida es Cardiotocography del respositorio de bases de datos UCI la cual la podemos descargar **aquí**. En esta base de datos se recogen distintas características de cardiotrogafías en las cuales se mide la frencuencia cardiaca fetal (FHR), los movimiento fetales (FM) y las contracciones uterinas (UC), obteniendo las siguientes características a partir de estos datos:

1. LB: punto de referencia del FHR en pulsaciones por minuto.
2. AC: aceleraciones del pulso por segundo.
3. FM: movimientos fetales por segundo.
4. UC: contracciones uterinas por segundo.
5. DL: deceleraciones suaves por segundo.
6. DS: deceleraciones fuertes por segundo.
7. DP: deceleraciones prolongadas por segundo.
8. ASTV: porcentaje de tiempo con variaciones anormales cortas del pulso.
9. MSTV: media de las variaciones anormales cortas del pulso.
10. ALTV: porcentaje de tiempo con variaciones anormales largas del pulso.
11. MLTV: media de las variaciones anormales largas del pulso.
12. Width: amplitud del histograma FHR.
13. Min: mínimo del histograma FHR.
14. Max: máximo del hisotograma FHR.
15. Nmax: número de picos en el histograma.
16. Nzeros: número de ceros en el histograma.
17. Mode: moda del histograma.
18. Mean: media del histograma.
19. Median: mediana del histograma.
20. Variance: varianza del histograma.
21. Tendency: tendencia del histograma.
22. CLASS: código del tipo de patrón del histograma FHR [1-10].

23. NSP: código del estado fetal. [1: Normal, 2: Sospechoso y 3: Patológico]

Lo que queremos es emplear estos datos para poder predecir ante una nueva cardiotocografía si el estado del feto es normal, sospecho o patológico, es decir, vamos a predecir la variable NSP con el resto.

```
datos <- read.csv("/media/griger/USB20FD/AA/datos.csv")
```

2. Codificación de los datos de entranda para hacerlos útiles a los algoritmos.

Nuestra base de datos estaba contenida en una hoja de cálculo. Para poder usarla dentro de R lo que hemos hecho es generar un CSV con los datos previamente formateados puesto que hemos tenido que cambiar el formato decimal de algunas columnas para que fuese el que emplea R. Además en el fichero original aparecían más variables como la fehca y el tiempo de inicio y fin de la cardiotocografía las cuales no hemos considerado relevantes para el estudio por lo que no están presententes en el CSV.

3. Valoración del interés de las variables para el problema y selección de un subconjunto en su caso.

En primer lugar tenemos que `Width` se calcula como la diferencia entre `Max` y `Min` con lo cual suponemos que una de las tres no tendrán relevancia ya que la información aportada por ella se puede deducir de las otras dos.

Para el resto de variables dado el poco conocimiento que tenemos en la materia no podemos saber qué factores son los que más influyen en determinar el estado del feto por tanto hemos decidido realizar un análisis de componentes principales para ver si podemos reducir el número de variables a considerar, haciendo que los algoritmos sean más eficientes en tiempo. La técnica que hemos usado en clase para tal propósito ha sido emplear el Lasso para obtener aquellas variables que sus coeficientes estuviesen por encima de un cierto umbral determinado por nosotros. Esto precisamente es lo que nos ha llevado a decantarnos por el PCA ya que con él podemos saber el conjunto de variables que son capaces de explicar al menos 95% de la variabilidad de los datos (aunque podemos cambiar este 95% y aumentarlo para que sea más estricto). Para saber cómo emplear PCA en R hemos consultado el enlace [2].

Bibliografía

1. La base de datos: https://archive.ics.uci.edu/ml/datasets/Cardiotocography#
2. PCA con R: http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/