APRENDIZAJE SUPERVISADO I

PROYECTO

Juan José Silva Torres

Introducción

Bienvenidos al proyecto del módulo *Análisis Supervisado I*. A lo largo del módulo se presentan los fundamentos de las etapas de un proyecto de modelización supervisada, desde el análisis del conjunto de datos inicial hasta la validación del algoritmo escogido. Como algoritmos de modelización supervisada se presentan en detalle ***Naive Bayes*** y ***Support Vector Machine* (SVM)**.

Para poner en práctica los conocimientos adquiridos durante las clases, el alumno debe trabajar semanalmente en un proyecto cuyo objetivo es ajustar un modelo, basado en el algoritmo de *Naive Bayes* y/o *SVM*, que ayude a los profesionales sanitarios a estimar el estado de salud fetal durante un parto. Para ello, se dispone de un conjunto de datos que contiene información relativa al estado de salud del feto y de su madre durante el parto. La mayoría de las variables explicativas se consiguen gracias a un sistema de monitorización que recoge de manera continuada las constantes vitales tanto del feto como las de la madre.

Tanto proyecto como *sprints* se deben implementar en Python, lenguaje de programación de propósito general muy popular, que además es ampliamente utilizado por analistas, ingenieros y científicos de datos. Como entorno de trabajo se escoge *Google Colab*, entorno que pone Google a disposición de la comunidad para que realicen sus desarrollos utilizando *Jyputer Notebooks*.

Sin duda es un proyecto muy atractivo que además de tener un alto valor ya que este tipo de algoritmos son muy susceptibles de implementarse en sistemas de producción que ayuden a los profesionales sanitarios a tomar mejores decisiones en momentos tan delicados como puede ser el nacimiento de un bebé.

Competencias y destrezas que adquirirá el alumn@

A modo de resumen, en el desarrollo proyecto el estudiante adquirirá las siguientes destrezas o competencias:

- Conocer y desarrollar las etapas a llevar a cabo en un proceso de modelización.
- Realizar el análisis descriptivo del conjunto de datos identificando la naturaleza de las variables explicativas y del *target* o evento a predecir.
- Creación de nuevas variables que pueden ser utilizadas en el proceso de modelización.
- Utilizar *Naive Bayes* como algoritmo de modelización supervisada.
- Utilizar *Support Vector Machine* como algoritmo de modelización supervisada
- Validar el modelo a partir de la matriz de confusión, la curva ROC y el área bajo la curva.
- Iniciación a *Google Colab* como entorno de trabajo.
- Iniciación a Python, concretamente a Pandas, Numpy y Scikit.

Entorno de trabajo

Se recomienda al alumno utilizar *Google Colab* como entorno de trabajo. *Google Colab* es una herramienta gratuita de Google, que te permite disfrutar de un entorno de pruebas basado en *Jupyter Notebook*, que te permite programar en Python y probar diferentes modelos. Ha cogido gran popularidad en los últimos años, ya que permite la utilización de una GPU, para entrenar modelos de *Machine Learning* y *Deep Learning* mucho más rápido que en un CPU.

Presenta las siguientes ventajas:

- No requiere configuración.
- Da acceso gratuito a GPUs.
- Permite compartir contenido fácilmente.

Referencias:

https://www.youtube.com/watch?v=inN8seMm7UI

Conjuntos de datos

Las complicaciones en el parto son una de las principales causas de mortalidad perinatal. El cardiotocógrafo fetal (CTG) se puede utilizar como una herramienta de seguimiento durante el parto ya que puede permite evaluar el estado de salud del bebé durante el nacimiento.

La cardiotocografía registra simultáneamente la frecuencia cardíaca fetal, los movimientos fetales y las contracciones uterinas. El registro permite al profesional sanitario valorar el latido cardíaco fetal durante la última etapa de la gestación y la respuesta del bebé a las contracciones durante todo el parto hasta el nacimiento.

Algunos términos importantes en el campo de la cardiotocografía:

- La traza CTG generalmente muestra dos líneas. La línea superior es un registro de la frecuencia cardíaca fetal en latidos por minuto. La línea inferior es un registro de las contracciones uterinas.

- Las cuatro funciones de frecuencia cardíaca fetal son la frecuencia cardíaca basal, variabilidad, aceleraciones y desaceleraciones.

- Las contracciones uterinas se cuantifican como el número de contracciones presentes en un período de 10 minutos y se promedian durante 30 minutos. Se suele considerar normal por debajo de 5 contracciones en 10 minutos y alto por encima de 5 contracciones en 10 minutos.

- La frecuencia cardíaca inicial es la frecuencia cardíaca fetal inicial promedio. Función tranquilizadora: 110-160 latidos por minuto (lpm). Función no tranquilizadora: 100-109 lpm o 161-180 lpm. Característica anormal por encima de 180 lpm.

- La variabilidad son las fluctuaciones en la frecuencia cardíaca fetal. Esto hace que el trazado aparezca como una línea irregular, en lugar de suave. La variabilidad es indicativa de un sistema neurológico fetal maduro y se considera una medida de la reserva fetal. Se considera característica tranquilizadora por encima de 5 lpm. Característica no tranquilizadora por

debajo de 5 lpm entre 40 minutos y 90 minutos. Característica anormal por debajo de 5 lpm durante más de 90 minutos.

- Las desaceleraciones son disminuciones en la frecuencia cardíaca fetal desde la línea de base en al menos 15 lpm, con una duración de al menos 15 segundos. Hay tres tipos de desaceleraciones, según su relación con la contracción uterina. La desaceleración temprana comienza al comienzo de la contracción uterina y termina con la conclusión de la contracción:

    - Característica tranquilizadora: sin desaceleración.
    - Característica no tranquilizadora: desaceleración temprana, desaceleración variable o desaceleración prolongada única hasta 3 minutos.
    - Característica anormal: desaceleraciones variables atípicas, desaceleración tardía o desaceleración prolongada única mayor de 3 minutos.

En este sentido, se recogieron datos de 2.126 cardiotocogramas fetales (CTG) y se midieron las características de diagnóstico respectivas. Los CTG también fueron clasificados por tres obstetras expertos y se asignó una etiqueta de clasificación de consenso a cada uno de ellos. De esta manera las características de diagnóstico representan el conjunto de variables explicativas que va a permitir estimar el estado fetal (target) en normal o anormal.

Sprint semanales y Entrega Final

Semana 1 – Análisis descriptivo

El estudiante debe familiarizarse con el conjunto de datos. Como primer paso es necesario realizar un análisis descriptivo que ayude al estudiante a comprender la naturaleza de las variables recogidas y del evento a predecir. Las cuestiones que debe responder son las siguientes:

- Clasificar cada variable según su naturaleza.
- ¿Cuál es la proporción de estados fetales normal? ¿y de anormales?
- Enumere 3 variables cuantitativas continuas junto con sus medidas de centralización, localización y dispersión.
- Obtenga el histograma para las variables continuas.
- Obtenga un gráfico *box -plot* para las variables continuas.
- Obtenga una distribución de frecuencias para las variables cualitativas o cuantitativas discretas con pocos valores.

Semana 2 – Modelización con Naive Bayes

El estudiante debe continuar con el proceso de modelización. En esta etapa se analizará la relación univariante del *target* con las variables explicativas, se avanza en la fase de muestreo creando un conjunto de entrenamiento y otro de test, para posteriormente ajustar un modelo con el algoritmo de *Naive Bayes*. Algunas tareas que debe realizar son las siguientes:

- Enumere las 3 variables explicativas que presentan mayor correlación de Pearson con el *target*.
- Cree un conjunto de entrenamiento y otro de test a partir del conjunto inicial de datos en el que el conjunto de entrenamiento sea el 60% de las observaciones totales.
- A partir del conjunto de entrenamiento entrene un modelo a partir del algoritmo de Naive Bayes. En función de la naturaleza de las variables explicativas escoja la familia que mejor se ajuste.
- Obtenga la curva ROC y el área bajo la curva para los conjuntos de entrenamiento y test.

- Evalúe si hay sobreajuste del modelo.

Semana 3 – Modelización con SVM

El estudiante debe continuar con el proceso de modelización. Se crea un nuevo conjunto de variables mediante la transformación z – score para posteriormente ajustar un modelo mediante el algoritmo *SVM*. Algunas cuestiones que debe responder son las siguientes:

- A partir del conjunto de entrenamiento y test creados en el segundo sprint, ajuste un modelo utilizando el algoritmo *SVM*.
- Obtenga la curva ROC y el área bajo la curva para los conjuntos de entrenamiento y test.
- Compare el área bajo la curva para el conjunto de test del modelo obtenido con *Naive Bayes* y *SVM*.

Semana 4 - Entrega final

En la última semana del módulo el alumno debe esforzarse en ajustar el mejor modelo posible. Una vez tenga el modelo, el alumno debe entregar un documento de un máximo de 3 folios que contenga los siguientes apartados:

- Análisis descriptivo de las variables explicativas y el *target*.
- Detalle de la división del conjunto de datos en los conjuntos de entrenamiento y test.
- Algoritmo seleccionado junto con los hiperparámetros escogidos en la modelización.
- Curva ROC y Área Bajo la Curva para los conjuntos de entrenamiento y test.

Además, el alumno deberá adjuntar el notebook final con el código desarrollado.