Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Facultad de Ingeniería

Maestría en Explotación de Datos y Descubrimiento de Conocimiento

Aprendizaje Automático

1er cuatrimestre de 2021

Trabajo práctico Nro 1

El objetivo de este trabajo práctico es analizar las particularidades de la utilización de algoritmos para la generación de árboles de decisión por medio de su aplicación en casos casi reales. El mismo pretende fijar conceptos estudiados en la teoría: sobreajuste y poda; tratamiento de datos faltantes; tolerancia al ruido; discretización de atributos numéricos. El material básico para la elaboración del presente trabajo se encuentra en las teóricas y prácticas presentadas hasta el momento y en las próximas clases y en la bibliografía indicada (por ej. libro de Mitchell). Podrá utilizarse cualquier otra fuente siempre que esté correctamente referenciada.

El presente trabajo será grupal. El grupo deberá estar compuesto por exactamente tres integrantes. Preferentemente uno de ellos debe saber programar. Se podrán evaluar contenidos del Trabajo Práctico durante el parcial posterior a la entrega del TP. Todos los integrantes deben tener conocimiento del desarrollo del TP.

La fecha límite de entrega es el 30 de mayo a las 23:59 hs.

Para el desarrollo del trabajo se utilizará un conjunto de datos que proveerá la cátedra (healthcare-dataset-stroke-data.csv) con el objetivo de predecir qué paciente va a sufrir un accidente cerebrovascular. Para resolver este problema se utilizarán árboles de decisión.

Se deberá elaborar un informe preferentemente en LaTeX y entregarlo en formato .pdf. La entrega deberá estar acompañada de la Jupyter Notebook en Python utilizada para generar los resultados. El documento a entregar debe cumplir con los siguientes requisitos:

- debe tener no más que cuatro hojas, con fuente tamaño 10 e interlineado simple. La bibliografía no cuenta en la cantidad de hojas.
- una carátula en donde figuren universidad, nombre de maestría, materia, número de grupo, nombres de los integrantes del grupo, número de TP, año de cursada, etc. La carátula no cuenta en la cantidad de hojas.
- un resumen (del estilo de un artículo científico de no más de 200 palabras)
- una introducción en donde, entre otros, conste el objetivo del trabajo y una explicación de cómo está organizado el resto del documento.
- una sección de datos, en donde se describan los datos utilizados y sus particularidades

una sección de metodología, en donde se describan las metodologías utilizadas (sobre datos y sobre algoritmos)

una sección resultados, que incluya los resultados y su análisis

una sección de conclusiones. Por tratarse de un trabajo de investigación netamente práctico, las conclusiones deben ser la resultante de la elaboración de las pruebas realizadas. La información obtenida de referencias externas puede y debe ser tomada como insumo, pero no como conclusión.

referencias bibliográficas (referenciadas a lo largo del trabajo)

El informe se deberá publicar en el aula virtual de la materia por uno sólo de los integrantes del grupo.

Para realizar el informe se deberán considerarse y documentarse los siguientes puntos:

a) A partir de los datos entregados, describir los atributos realizando una breve explicación de qué representan y del tipo de variable (categórica, numérica u ordinal). En caso de que haya variables no numéricas, reportar los posibles valores que toman y cuán frecuentemente lo hacen.

b) Reportar si hay valores faltantes. ¿Cuántos son y en qué atributos se encuentran? En caso de haberlos, ¿es necesario y posible asignarles un valor?

c) ¿Qué variables se correlacionan más con el evento de lesión (Stroke)? Para las cuatro más correlacionadas, realizar un gráfico en el que se pueda observar la correlación entre la variable y el stroke.

d) Se necesita saber cuáles son los indicadores que determinan más susceptibilidad a sufrir una lesión. ¿Qué atributos utilizará como variables predictoras? ¿Por qué?

e) ¿Se encuentra balanceado el conjunto de datos que utilizará para desarrollar el algoritmo diseñado para contestar el punto d)? En base a lo respondido, ¿qué métricas de performance reportaría y por qué? En caso de estar desbalanceado, ¿qué estrategia de balanceo utilizaría?

f) Suponiendo que es más importante detectar los casos en donde el evento ocurre. ¿Qué medida de performance utilizaría? Si utiliza Fβ-Score, ¿qué valor de β eligiría?

g) Implementar el algoritmo introducido en el punto d) utilizando árboles de decisión. En primer lugar, se deberá separar un 20% de los datos para usarlos como conjunto de evaluación (test set). El conjunto restante (80%) es el de desarrollo y es con el que se deberá continuar haciendo el trabajo. Realizar los siguientes puntos:

1) Armar conjuntos de entrenamiento y validación con proporción 80-20 del conjunto de desarrollo de forma aleatoria. Usar 50 semillas distintas y realizar un gráfico de caja y bigotes que muestre cómo varía la métrica elegida en c) en esas 50 particiones distintas.

2) Usar validación cruzada de 50 iteraciones (50-fold cross validation). Realizar un gráfico de caja y bigotes que muestre cómo varía la métrica elegida en esas 50 particiones distintas.

h) Graficar el árbol de decisión con mejor performance encontrado en el punto g2). Analizar el árbol de decisión armado (atributos elegidos y decisiones evaluadas).

i) Usando validación cruzada de 10 iteraciones (10-fold cross validation), probar distintos valores de α del algoritmo de poda mínima de complejidad de costos (algoritmo de poda de sklearn). Hacer gráficos de la performance en validación y entrenamiento en función del α. Explicar cómo varía la profundidad de los árboles al realizar la poda con distintos valores de α.

j) Evaluar en el conjunto de evaluación, el árbol correspondiente al α que maximice la performance en el conjunto de validación. Comparar con el caso sin poda (α=0)

k) Para el árbol sin poda, obtener la importancia de los descriptores usando la técnica de eliminación recursiva. Reentrenar el árbol usando sólo los 3 descriptores más importantes. Comparar la performance en el conjunto de prueba.

Atención: los puntos anteriores no necesariamente deben ser respondidos en el mismo orden en el que son formulados. El único requisito es que sus respuestas estén en alguna parte del informe entregado.