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 2019

Trabajo práctico Nro 1

El objetivo de este trabajo práctico es construir clasificadores basados en árboles de decisión y de Naive Bayes para casos cuasi reales.

Se deberá obtener un data set de no menos de 20 atributos y 500 ejemplos clasificados en no más de tres clases. De los 20 atributos al menos 4 y no más de 8 deberán ser numéricos. El resto deberán ser categóricos de no más de 5 clases cada uno. Es preferible que los datos correspondan a un problema real. De no conseguirse, se podrán utilizar datasets públicos. En cualquier caso, se deberá corroborar con los docentes auxiliares que los datos son adecuados.

Ejercicios

1. Partición de datos

Particionar el conjunto de datos en entrenamiento, validación y test. El conjunto de test se deberá dejar apartado para ser utilizado al final.

2. Árboles de decisión

1. **Entrenar un árbol de decisión con altura 3** y el resto de los hiperparámetros con su valor en default. Estimar la performance del modelo utilizando **5-fold cross validation** utilizando el **Accuracy** y **ROC AUC**. Informar:
    - para cada fold el **Accuracy** y **ROC AUC** para
        - Conjunto de entrenamiento
        - Conjunto de validación
        - Promedio y desviación estándar de:
            - todos los conjuntos de entrenamiento
            - todos los conjuntos de validación
2. **Entrenar árboles de decisión con las siguientes combinaciones**. En todos los casos probar e informar **Accuracy** y **ROC AUC** para training y para validación con Gini y con Information Gain haciendo **cross validation**:
    a. Altura máxima 3
    b. Altura máxima 6
    c. Sin límite de altura máxima

3. **Tratamiento de datos faltantes.** Probar las siguientes alternativas para completar los datos faltantes:

- Moda: se rellena el dato faltante con la moda del atributo.
- Moda de clase: se rellena el dato faltante con la moda del atributo según la clase.

La función para implementación de datos faltantes deberá tomar como parámetros de entrada el dataset, el porcentaje de faltantes y la estrategia de relleno. La salida generará los datos de entrada con los datos faltantes rellenados correspondientemente. Si el dataset utilizado no contiene datos faltantes, generarlos adrede.

Construir una familia de datasets agregando datos faltantes para cada estrategia de relleno sobre el 80% de los datos desarrollo. Preservar un 20% sin alterar para las corridas de validación. Variar desde 0% a 80% en intervalos de 5%.

a) Ejecutar corridas del mejor método del punto 2 para cada familia.
b) Graficar el tamaño del árbol en función del porcentaje de faltantes.
c) Graficar la performance (en Accuracy) en función del porcentaje de faltantes. Utilizar el 20% de datos inalterados para validación.
d) Analizar el tratamiento de datos faltantes de las distintas estrategias sobre los resultados obtenidos.

4. **Tolerancia al ruido.** Se deberá implementar una función que introduzca ruido (cambio de un valor) sobre un atributo numérico.

Para este experimento se deberá implementar una función ind_ruido, que toma como parámetros de entrada el dataset y el porcentaje de ruido a introducir sobre la clase

a) Construir una familia de datasets generando ruido. Preservar un 20% sin alterar para las corridas de validación. Variar desde 0% a 35% en intervalos de 5% el atributo elegido.
b) Ejecutar corridas del mejor método del punto 2 para cada familia.
c) Graficar el tamaño del árbol en función del porcentaje de ruido.
d) Graficar la performance en función del porcentaje de ruido. Utilizar el 20% de datos inalterados para validación.
e) Analizar la tolerancia al ruido sobre los resultados obtenidos.

3. Naive Bayes

Ejecutar Naive Bayes, informar las probabilidades condicionales y previas. Realizar las validaciones correspondientes

4. Comparación de algoritmos

Comparar Naive Bayes y árboles de decisión. Para hacerlo usar 5-fold cross-validation para la exploración de la mejor solución en cada caso. En árboles de decisión determinar qué tamaño de árbol conviene y si conviene utilizar Gini o Information Gain (todo esto con el conjunto de desarrollo y utilizando grid search).

Utilizar ROC AUC como métrica.

Testear ambos algoritmos con el conjunto de test independiente (mismo conjunto de test en ambos casos).

Otros detalles

El grupo deberá estar compuesto por exactamente tres integrantes. Preferentemente uno de ellos debe saber programar.

Se podrán evaluar contenidos del Trabajo Práctico durante los parciales posteriores a la entrega del TP. Todos los integrantes deben tener conocimiento del desarrollo del TP.

El trabajo deberá implementarse en Python y entregado en una Jupyter notebook con el correspondiente data set para poder ser probado.

Informe

El documento a entregar debe cumplir con los siguientes requisitos:

- una carátula en donde esté el nro. del grupo, sus integrantes, nombre de maestría, materia, etc.

- un resumen (del estilo de un artículo científico)

- 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 metodologías, 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 y trabajos futuros. 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 no deberá tener más de 10 páginas a espacio simple en Arial 11 y se deberá publicar en el aula virtual de la materia por uno sólo de los integrantes del grupo. También deberá entregarse impreso.

Fecha de entrega: 20 de mayo