Pontificia Universidad Católica de Chile
Educación Profesional - Escuela de Ingeniería
Diplomado en Big Data y Ciencias de Datos
Minería de Datos
Relator: Sebastián Raveau

Evaluación 2 – Métodos de Clasificación

Fecha de Entrega: domingo 30 de julio

El objetivo de esta evaluación es utilizar, analizar y comparar dos algoritmos de clasificación: árboles de clasificación y KNN. La base de datos contiene registros de 1.000 de evaluación de riesgo crediticio de un banco alemán. La descripción de las variables de la base de datos se encuentra en la siguiente página.

La evaluación consta de cuatro etapas:

1. Dividir la base de datos en dos, una base de entrenamiento (mayoritaria) y una base de validación (minoritaria), explicando el proceso.
2. Aplicar árboles de clasificación, analizando el impacto en los resultados de: i) el método de división: *information* vs *gini*; ii) del parámetro *cp* asociado al costo de complejidad; y iii) los parámetros de parada anticipada *minsplit*, *minbucket* y *maxdepth*. Cada comparación es independiente entre sí (y por lo tanto debe realizar tres comparaciones). La comparación consiste en una evaluación crítica del árbol resultante (e.g. profundidad, variables utilizadas, divisiones realizadas, etc.) y la capacidad predictiva del modelo (obteniendo, por ejemplo, matrices de confusión).
3. Aplicar KNN, analizando el parámetro K del algoritmo y las variables a utilizar para la clasificación (en función de su relevancia y conveniencia). Tras el análisis deberá proponer la especificación que le parezca más adecuada, usando una o más métricas de rendimiento que le parezcan adecuadas (por ejemplo, comparar *acccuracy*).
4. Comparar brevemente los resultados de ambos algoritmos.

Debe generar un breve reporte y entregarlo al correo: mineria.datos.PUC@gmail.com. El asunto del correo y el nombre del archivo deben comenzar con [Evaluación 2] seguido con los apellidos de los estudiantes. Por ejemplo: [Evaluación 2] Gutiérrez y Soto. El objetivo de la evaluación es demostrar que es capaz de aplicar los contenidos del curso. Por favor presente sólo la información relevante, sin llenar múltiples páginas con gráficos, códigos y estadísticas. Extensión recomendada: 10 planas de contenido (i.e. sin contar portada, índices o anexos, los cuales no son obligatorios).

Pontificia Universidad Católica de Chile
Educación Profesional - Escuela de Ingeniería
Diplomado en Big Data y Ciencias de Datos
Minería de Datos
Relator: Sebastián Raveau

Descripción de la Base de Datos (según la fuente original)

Nombre	Descripción	Tipo	Valores
status	Historial de cuenta corriente	Categórica	1 = sin cuenta 2 = cuenta con deuda 3 = ahorros hasta 200 DM 4 = ahorros desde 200 DM
history	Historial de créditos	Categórica	0 = retrasos en el pasado 1 = cuenta crítica o créditos en otros bancos 2 = sin créditos anteriores 3 = créditos vigentes al día 4 = todos los créditos pagados
amount	Monto del crédito (transformado por confidencialidad)	Continua	-
savings	Historial de ahorros	Ordinal	1 = Sin ahorros 2 = Menos de 100 DM 3 = Entre 100 y 500 DM 4 = Entre 500 y 1.000 DM 5 = Más de 1.000 DM
employed	Tiempo empleado	Ordinal	1 = Desempleado 2 = Menos de 1 año 3 = Entre 1 y 4 años 4 = Entre 4 y 7 años 5 = Más de 7 años
rate	Tasa del crédito en proporción del ingreso	Ordinal	1 = Más de 35% 2 = Entre 25% y 35% 3 = Entre 20% y 25% 4 = Menos de 20%
personal	Estado civil y sexo	Categórica	1 = Hombre divorciado 2 = Hombre soltero * 2 = Mujer no soltera * 3 = Hombre casado o viudo 4 = Mujer soltera
residence	Tiempo en la residencia	Ordinal	1 = Menos de 1 año 2 = Entre 1 y 4 años 3 = Entre 4 y 7 años 4 = Más de 7 años
property	Bien más valioso	Ordinal	1 = Nada 2 = Auto 3 = Seguro de vida 4 = Propiedad
age	Edad	Continua	-

housing	Tipo de vivienda	Categórica	1 = Gratuita 2 = Arrendada 3 = Propia
credits	Créditos vigentes	Ordinal	1 = 1 2 = 2 o 3 3 = 4 o 5 4 = 6 o más
job	Nivel de empleabilidad	Ordinal	1 = desempleado 2 = no calificado 3 = calificado dependiente 4 = gerente / independiente
persons	Carga familiar	Binaria	0 = 2 o menos , 1 = 3 o más
telephone	Tiene teléfono fijo	Binaria	0 = no , 1 = sí
foreign	Extranjero	Binaria	0 = no , 1 = sí
credit	Evaluación de riesgo	Binaria	0 = mala , 1 = buena

Su entrega será evaluada de acuerdo con los siguientes aspectos:

Claridad [1 punto]

Se espera que el reporte sea claro y bien redactado, con las ideas debidamente desarrolladas. Las tablas y figuras deben estar debidamente presentadas y explicadas. El procedimiento de aplicación del algoritmo se explica adecuadamente. El lenguaje utilizado debe ser apropiado para un reporte técnico.

Aplicación de árboles de decisión [2 puntos]

Se presentan y comparan los resultados de distintos métodos de división, parámetros de complejidad y parámetros de parada anticipada. Los parámetros utilizados se presentan y fundamentan en forma explícita. Se discuten los resultados desde una perspectiva crítica (e.g. profundidad, variables utilizadas, divisiones realizadas, etc.). Se presenta y compara la capacidad predictiva de los árboles resultantes.

Aplicación de KNN [2 puntos]

El algoritmo se debe aplicar correctamente. Se debe explicitar y justificar la selección del parámetro K. Se deben explicitar y justificar las variables utilizadas en el modelo. Se debe explicitar y justificar la especificación que le parezca más adecuada, en función de las métricas de rendimiento utilizadas

Comparación [1 punto]

Se comparan y discuten los resultados de ambos algoritmos de clasificación. Las métricas de rendimiento deben ser apropiadas y utilizadas adecuadamente.