PD2

3

El workflow importa el dataset iris, lo divide en una parte de entrenamiento y otra de testeo (80% y 20%) usando sampleo estratificado que mantiene la proporción del dataset original. Luego se entrena un modelo de árbol de regresión simple con el ancho de pétalo como columna objetivo, luego aplica el modelo con los datos de testeo. Finalmente muestra una comparación entre el valor verdadero y el valor de predicción en una grafica para mostrar la performance del modelo.

4

En rapid miner primero se hace la importación de los datos, luego se coloca el operador que devuelve el dataset y luego se setea el rol con otro operador, en kmine está todo junto en el operador file reader. Los números están representados como double y la clase como string.

La variable de predicción se define en el operador tree learner.

5

El particionado está hecho para solo hacer dos particiones, en rapid minner se pueden hacer múltiples. RM da la opción de que la forma en la que se hace el Split se elija de forma automática mientras que en kamine hay que setearla. Kmine da la opción que elegir que la partición tenga n valores y el resto de los valores sean la otra partición.

6

Soporta solo variables de predicción númericas, los predictores pueden ser de numéricos o nominales.

El algoritmo usado sigue el procedimiento descripto por Breiman pero la implementación del operador tiene algunas simplificaciones. En un árbol de regresión el valor predicho para un nodo hoja es el valor objetivo medio de los registros dentro de la hoja. El algoritmo trata de que la varianza en los registros de una hoja sea mínima, para ello minimiza la suma de los errores cuadrados de los hijos cuando va a hacer una división.

Los parámetros para configurar son:

Use binary splits for nominal atribute: Si se selecciona, el nodo determinará divisiones binarias basadas en conjuntos para los valores nominales. Si no se hace esto cada valor resultará en un nodo hijo.

Missing value handling: Este parametro presenta dos opciones:

- XGBoost: Si se selecciona esta opción (también es la predeterminada), el predictor calculará qué dirección es la más adecuada para los valores faltantes, enviando los valores faltantes en cada dirección de una división. La dirección que produce el mejor resultado (es decir, la mayor ganancia) se utiliza como dirección predeterminada para los valores perdidos. Este método funciona con divisiones binarias y multivía.

- Surrogate: Este método calcula para cada división divisiones alternativas que se aproximan más a la mejor. Solo puede ser usado con divisions nominales binarias.

Limit Number of leves: Pone límite a la profundiad del árbol.

Minimum Split node size: Número mínimo de valores en un nodo del árbol para que se haya intentado una división.

Minimum node size: Número mínimo de valores en los nodos secundarios. Puede ser como máximo la mitad del mínimum Split node size. Este parámetro se ignora para las divisiones nominales.

7

Acepta el modelo creado y los datos para predecir. El operador permite poner el nombre de la columna predicha.

8

Es una gráfica de puntos que tiene los valores verdaderos y predichos del largo de los pétalos. Los parámetros son el número de filas a graficar y el número de valores nominales que debe tener una columna para que sea ignorada. Los parámetros son bastante razonables porque ayudan a que la gráfica sea legible.

9

El numeric scorer compara los resultados predichos con las medidas reales, calcula el error, error cuadrado, sus promedios y todas las variables relacionadas.

10. Alterar el modelo para hacer el mismo tipo de análisis que en el Ejercicio 1 (comparar performance de training vs. Test, a. para distintos valores de los parámetros del AD, registrar los rendimientos (exactitud) obtenidos.

10

	4	10	15	20
2	1.168	0.17	0.148	0.141
4		0.156	0.15	0.141
6			0.123	0.125
8				0.123

En el eje vertical es mínimum node size y el horizontal es el mínimum Split node size. En la tabla están los valores del error absoluto promedio en las predicciones.