Diseño de un experimento con Impala

Miguel Ángel Roldán Carmona

1. Estructura y contenido de la base de datos

Para el diseño del experimento se ha escogido el conjunto de datos Wine Quality disponible aquí.

Este conjunto de datos contiene muestras de vino blanco del vino portugués "Vinho verde". El conjunto tiene 4899 instancias, 12 atributos y no presenta valores perdidos .

Es un conjunto muy versátil debido a que se puede utilizar para tareas de regresión y clasificación. Las clases están ordenadas y no balanceadas (e.g. hay más vinos normales que excelentes o pobres). Se pueden aplicar algoritmos de detección de outliers para detectar vinos excelentes o pobres. Además, no se conoce la importancia de las variables, por lo que podría ser interesante probar métodos para selección de características.

1.1. Información sobre los atributos

Variables de entrada (todas son numéricas):

- fixed acidity.
- volatile acidity.
- citric acid.
- residual sugar.
- chlorides.
- free sulfur dioxide.
- total sulfur dioxide.
- density.
- pH.

- sulphates.
- alcohol.

Variable de salida (numérica):

- quality (puntuación entre 0 y 10)

2. Consultas planteadas

2.1. Carga de ficheros

Accedemos a la shell y cargamos el fichero winequality-white.csv al sistema de ficheros hdfs para utilizarlo a continuación.

```
sudo bash
su - impala
ls /var/tmp/materialImpala
# mkdir /var/tmp/materialImpala
wget https://archive.ics.uci.edu/ml/machine-learning-databases/
    wine-quality/winequality-white.csv
cp winequality-white.csv /var/tmp/materialImpala

hdfs dfs -ls /user/impala/input
# hdfs dfs -mkdir /user/impala/input
hdfs dfs -put /var/tmp/materialImpala/winequality-white.csv
    /user/impala/input
hdfs dfs -ls /user/impala/input
```

2.2. Creación de la tabla

Una vez dentro de impala, creamos la tabla correspondiente. Hay que prestar atención al delimitador para cada campo (es un punto y coma, no una coma) y cuando se carga la tabla (se ignora la primera columna porque es la cabecera).

```
impala-shell
CREATE TABLE IF NOT EXISTS WineQuality (FixedAcidity FLOAT,
    VolatileAcidity FLOAT, CitricAcid FLOAT, ResidualSugar FLOAT,
    Chlorides FLOAT, FreeSulfurDioxide FLOAT, TotalSulfurDioxide FLOAT,
    Density FLOAT, Ph FLOAT, Sulphates FLOAT, Alcohol FLOAT, Quality INT)
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ';' STORED AS TEXTFILE;
```

```
DESCRIBE WineQuality;
LOAD DATA INPATH '/user/impala/input/winequality-white.csv'
    OVERWRITE INTO TABLE WineQuality IGNORE 1 LINES;
```

2.3. Consultas planteadas

Primero se considera utilizar una consulta básica para comprobar que los datos se han leído bien:

```
SELECT Quality FROM WineQuality;
```

Finalmente, se carga la consulta que selecciona 5 atributos y realiza una una operación de selección más compleja que involucre un operador AND (u OR) combinado con un operador OR (o AND) o con otro operador lógico distinto.

```
SELECT FixedAcidity, VolatileAcidity, Ph, Alcohol, Quality
FROM WineQuality WHERE (Quality > 6) AND ((Alcohol > 12) OR (Ph > 3));
```