Aprendizaje computacional. UNET Agosto 2012.

Profesores: Jose López y Jacinto Dávila

Actividad SVM

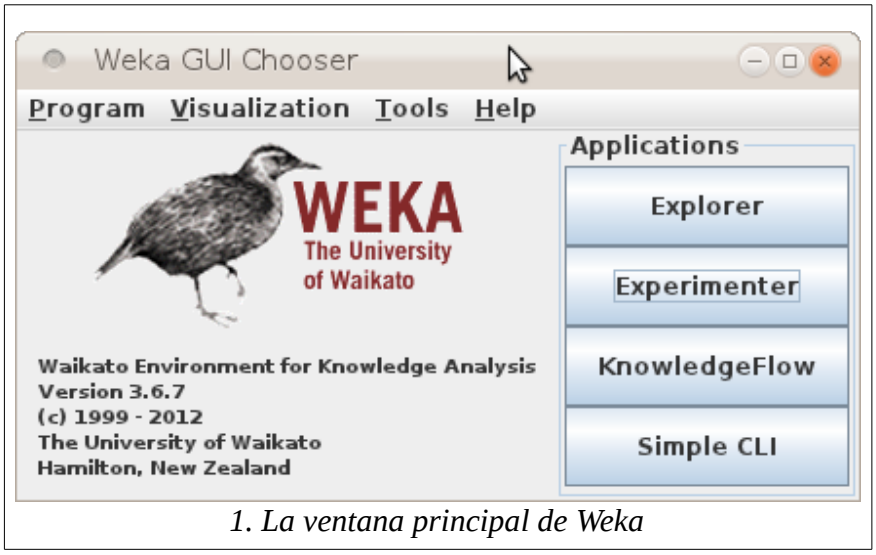
Objetivo: Generar modelos de regresión basados en máquinas de soporte vectorial para predecir parámetros en una mezcla óptima de gasolina a partir de componentes básicos industriales. La actividad está inspirada por el trabajo del Ing César Pernalete, "*Un agente inteligente para la optimización en línea del proceso de mezclas de gasolina* ", a ser presentado como tesis de la maestría en Modelado y Simulación de la Universidad de Los Andes, en el 2012. La data corresponde a una pequeña selección de los registros que preserva PDVSA Intevep.

Aún cuando el problema está inspirado por la tesis de Pernalete, la actividad se realizará con la herramienta WEKA que ha sido presentada en el curso, con la intención de que sirva como referencia y validación independiente de la estrategia y les ahorre a los estudiantes todo el esfuerzo de configuración y curva de aprendizaje de la plataforma R, usada en esa tesis.

La Actividad se divide en tres subactividades:

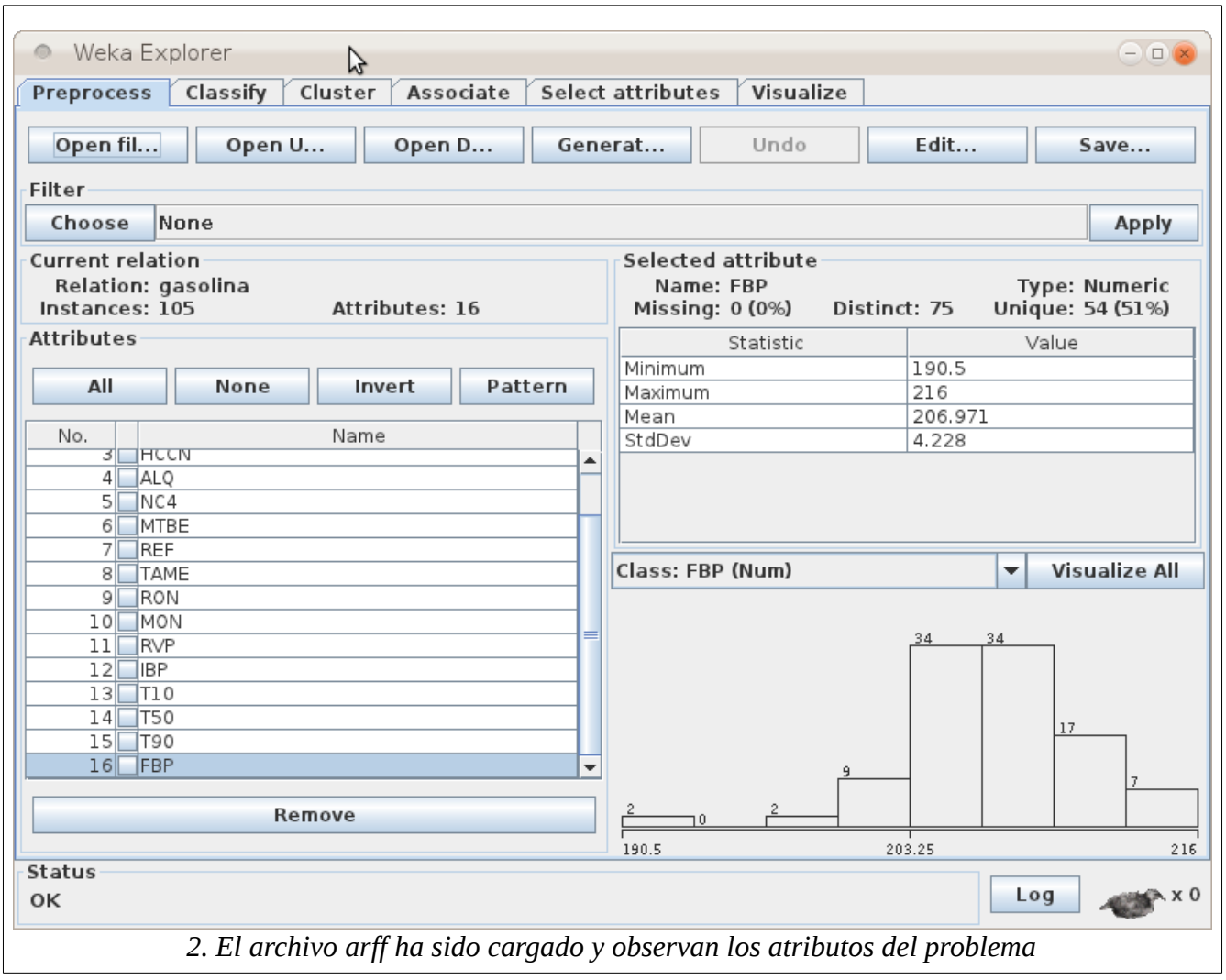
Subactividad 1: Validación de formatos y ejecución de clasificador lineal SMO de Weka.

Para esto, los estudiantes deberán procesar el archivo 91vPrueba.arff que se anexa. Ese archivo contiene la especificación de un problema de aprendizaje en los términos que requiere Weka. La siguiente es la secuencia de pantallas al cargar el archivo:



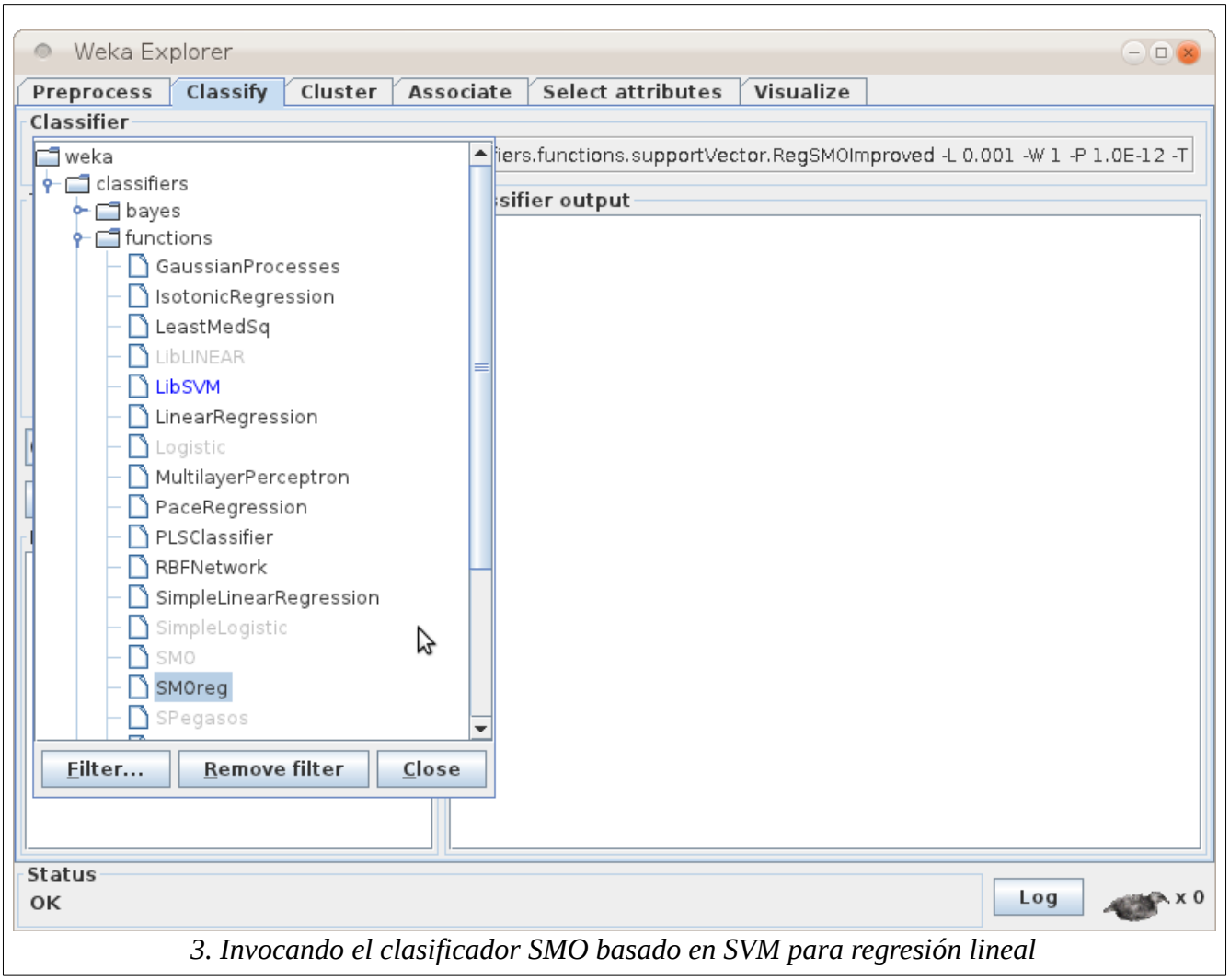
1. La ventana principal de Weka

En esa pantalla principal se debe abrir, con un clic, el Explorer y desde allí, abrir el archivo (con el primer botón en la esquina superior izquierda). Al cargar el archivo, la ventana lucirá así:



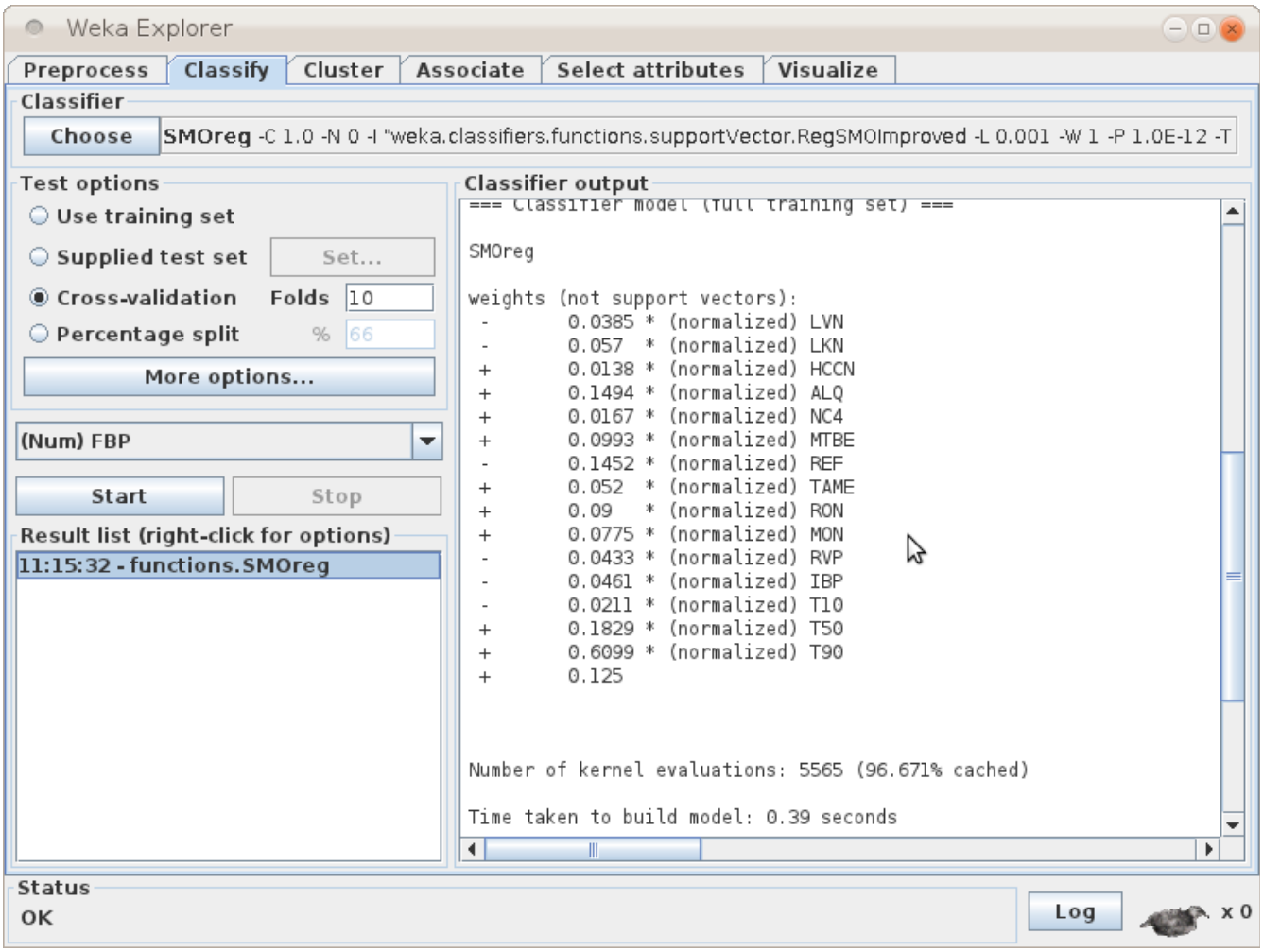
2. El archivo arff ha sido cargado y observan los atributos del problema

La subactividad culmina con la invocación de un algoritmo SVM de clasificación lineal intercontruído en el Weka. El SMO se puede invocar desde la interfaz gráfica Weka a través del tabulador *Clasify*, como se muestra a continuación:



3. Invocando el clasificador SMO basado en SVM para regresión lineal

Una vez seleccionado el algoritmo de clasificación (noten que Weka sólo permite algoritmos de regresión dados los tipos de los datos en atributo de clase del archivo arff), se debe iniciar el proceso de aprendizaje o generación del modelo predictor, con el botón *Start*. La salida lucirá así:



4. Illustration: Salida del SMOreg para el mezclador de gasolinas

En esta salida de la figura 4 se puede observar el conjunto de pesos que corresponden a cada atributo en el modelo de regresión basado en una máquina de soporte vectorial con un kernel lineal.

El equipo procederá a **interpretar** esas salidas (es probable que necesite re-ejecutar el algoritmo, cambiando algunos parámetros de salida con el botón "*more options*").

Subactividad 2: Una interpretación más realista del predictor de la mezcla de gasolina.

La primer subactividad es un ejercicio simplista extremo en el caso del predictor de gasolina. Se han colocado en el archivo arff todos los datos de todos los atributos que suelen almacenarse en una tabla con el registro de todas las prediciones para una mezclar y se ha declarado que el último atributo, RBF es el indicador de clasificación.

La situación real es más compleja. De hecho, los atributos independiente de la mezcla son solamente los que Pernalete explica a continuación:

```
Volumen total de mezcla: 100000 bbl
Inicialmente se asume una disponibilidad en tanques de:
LVN:  22000 bbl
LKN:  18000 bbl
HCCN: 110000 bbl
ALQ:  120000 bbl
nC4:  20000 bbl
MTBE: 130000 bbl
REF:  20000 bbl
TAME: 10000 bbl
------------------------
Los volúmenes a mezcla por componentes son:
LVN:  0 bbl
LKN:  15000 bbl
HCCN: 25000 bbl
ALQ:  0 bbl
nC4:  10000 bbl
MTBE: 40000 bbl
REF:  10000 bbl
TAME: 0 bbl
```

Los demás atributos corresponden a características de la calidad deseada en el producto final y que son el objeto de la especificación contractual (por lo que pagan los clientes de PDVSA). Pernalete sigue explicando:

```
Las restricciones de calidad en la mezcla final son:
RON >= 91.6
RVP <= 9.5
IBP >= 30
T10 =< 70
T50 >= 77
T50 =< 121
T90 =< 195
FBP =< 225
IAD =< 87.6
IAD >= 87
```

Para efectos de interpretación es también muy importante notar que los atributos independientes se expresan como un porcentaje de la disponibilidad en tanques. Así explica Pernalete:

```
Optimización a ejecutarse bajo las siguientes condiciones iniciales:
---------------------------------------
Porcentajes volumétricos de la mezcla:
LVN  -> 0 %
LKN  -> 15 %
HCCN -> 25 %
```

```
ALQ  -> 0 %
nC4  -> 10 %
MTBE -> 40 %
REF  -> 10 %
TAME -> 0 %
----------------------------------------
Límites inferiores en las proporciones de mezcla
LVN  -> 0 %
LKN  -> 0 %
HCCN -> 0 %
ALQ  -> 0 %
nC4  -> 0 %
MTBE -> 0 %
REF  -> 4 %
TAME -> 0 %
----------------------------------------
Límites superiores en las proporciones de mezcla
LVN  -> 22 %
LKN  -> 18 %
HCCN -> 100 %
ALQ  -> 100 %
nC4  -> 20 %
MTBE -> 100 %
REF  -> 20 %
TAME -> 10 %
----------------------------------------
```

Con lo cual luego se puede expresar los resultados del aprendizaje de los modelos SVM y la posterior tarea de optimización de la siguiente manera:

```
Número de iteraciones: 4413
**********************
La receta óptima es:
**********************
x0 LVN:  20.45532989
x1 LKN:  17.1774173
x2 HCCN: 34.36725282
x3 ALQ:  0.1841394953
x4 nC4:  2.026987138
x5 MTBE: 18.68274217
x6 REF:  6.981562047
x7 TAME: 0.1245691475
Costo de la mezcla: 54.61419545
------------------------
Las calidades obtenidas en la mezcla óptima son:
RON = 92.62017552
IAD = 87.37377465
RVP = 9.5
```

```
IBP = 34.87589812
T10 = 52.59023793
T50 = 82.30987981
T90 = 167.849415
FBP = 205.893396
------------------------
Tiempo de ejecución de la optimización 552 s
```

En esta subactividad 2, el equipo deberá organizar la data en el archivo arff para generar modelos **PARA CADA UNO** de los atributos de calidad por separado, vale decir:

```
RON
IAD
RVP
IBP
T10
T50
T90
FBP
```

Los modelos deberán ser usado, como se hizo en la actividad 2 (*more options*) para **predecir cada uno de esos atributos sobre la data dada, por separado** y el trabajo incluirá la **documentación** del ejercicio y un **análisis** de la capacidad predictiva de cada modelo (de cada atributo) para responder a la pregunta: *¿Cuál de los atributos independientes es más influyente en la calidad de la mezcla final?*

Subactividad 3: LibSVM

En esta actividad, el equipo repetirá el proceso de aprendizaje de la subactividad 2 sobre ese mismo archivo .arff, pero luego de **instalar y configurar la librería weka LibSVM**, lo cual les permitirá emplear otros kernels para generar modelos predictores no lineales.

El equipo **documentará** cada paso del nuevo proceso, así como las salidas correspondientes.

Subactividad 4: Weka desde Java

En esta actividad, el equipo automatizará las tareas de preparación y procesamiento de la data usada en las subactividad 1 y 2, por medio de un pequeño programa Java que invoque a Weka. Los detalles de integración de la librería o biblioteca Weka con Java se describen acá:

http://weka.wikispaces.com/Use+WEKA+in+your+Java+code

Particularmente, se espera que el programa Java contenga las instrucciones apropiadas para leer un archivo ARFF (**ARFF File Instance**), filtrar los atributos necesarios de esa data **(Option handling y Filtering on-the-fly),** invocar el clasificador SVM **(Classification),** realizar una validación cruzada **(Cross-validation)** y mostrar algunas estadísticas resultantes **(Statistics).**