Fundamentos de Ingeniería de Datos

Equipo 4

Introducción a BigML

Profundizando en OptiML

Autores:

Javier Ortiz Pérez

José Enrique Sánchez

Tabla de contenido

Introducción

El presente documento tiene como objetivo servir de breve introducción a la herramienta de Machine Learning BigML así como profundizar en las características fundamentales de una de sus funciones más importantes, OptiML.

Recursos

Cualquier cosa que pueda ser creada por parte del usuario en BigML es denominada Recurso. Existen dos tipos de recursos fundamentales, las Sources y los Datasets.

Vistas

Las vistas son la forma de organizar de forma visual los recursos creados por el usuario. Existen dos tipos fundamentales

- List view: Que permiten al usuario visualizar todos los recursos de un mismo tipo
- Resource view: Que permiten al usuario entrar en los detalles específicos de un recurso en concreto.

Adicionalmente, los recursos pueden ser almacenados en una estructura de carpetas que la herramienta denomina Projects

Sources

Son los recursos que permiten cargar datos en BigML.

Tipos de orígenes

La única forma de cargar datos en BigML es a través de ficheros (csv, tsv o arff)

Las formas de cargar dichos ficheros en la plataforma, sin embargo, son variadas:

- Carga Directa
- URLs
- Cloud Integration (Google/Dropbox)
- API / Bindings.

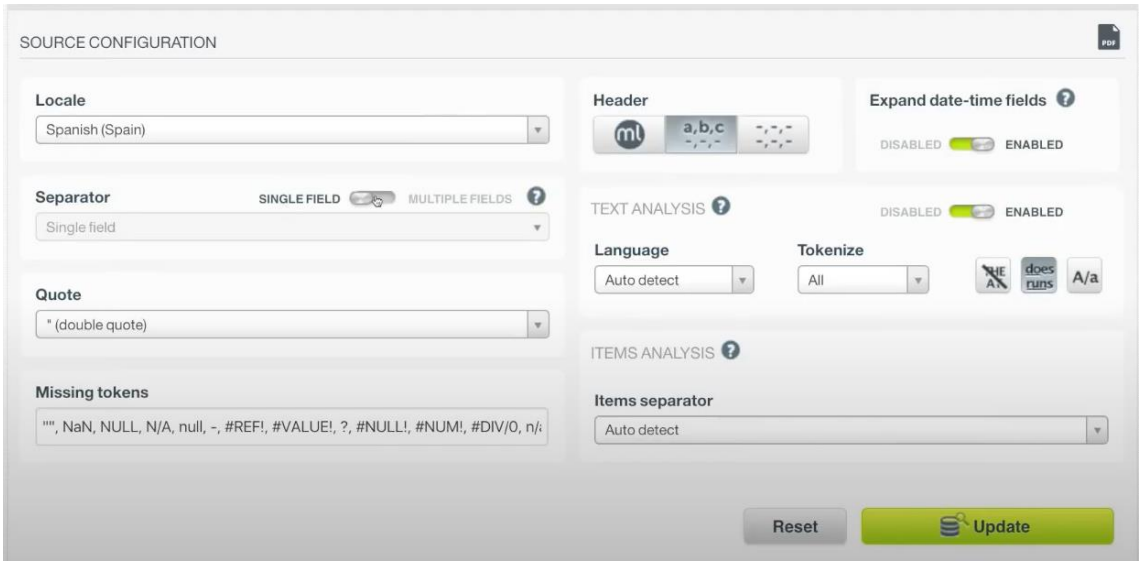
Configuración

Los sources ponen a nuestra disposición una serie de opciones que nos permitirán cargar los datos en el formato que nosotros deseemos.

Por ejemplo, Incluyen un detector automático del tipo de dato (numérico, categórico, etc.). Si bien, es importante revisar el trabajo automático realizado por el source y modificarlo manualmente si fuera necesario.

Por otra parte, hay una serie de configuraciones manuales como *Locale*, separadores o *header,* además de todas las especificaciones típicas que posee cualquier herramienta

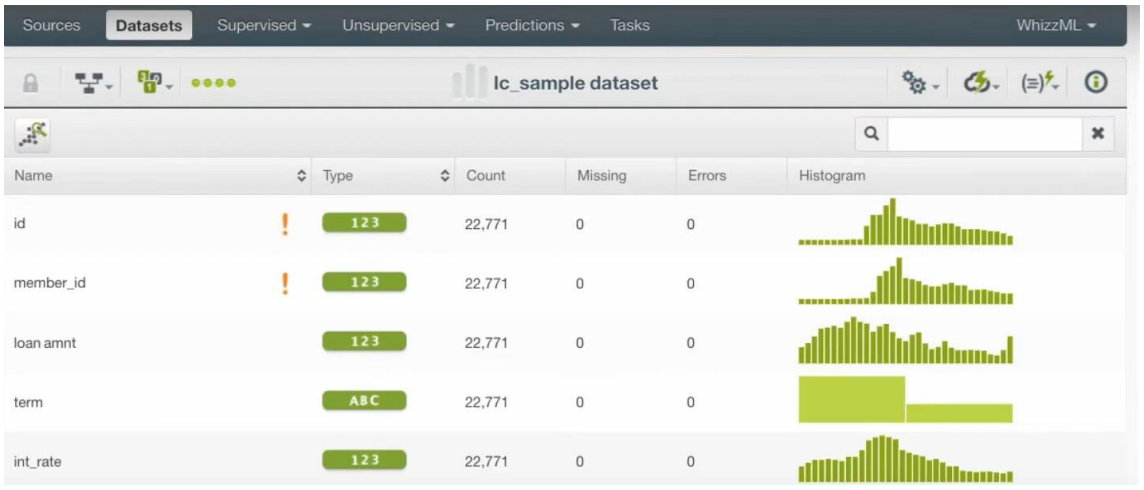
ETL sencilla. En la siguiente imagen, podremos ver la interfaz que proporciona la herramienta para editar las configuraciones que acabamos de mencionar.



Datasets

Son los recursos claves para la construcción de modelos en BigML y son creados a partir de los Sources.

Nos permiten una visión más profunda de los datos, proporcionándonos algunas estadísticas como las que podemos ver en la imagen inferior.



También nos permitirán llevar a cabo una serie de tareas previas al entrenamiento de los modelos, como, por ejemplo, la selección de atributos. La propia herramienta descartará automáticamente aquellos atributos que considera inútiles para el aprendizaje del modelo.

A continuación, proporcionamos un listado de las distintas funciones que se pueden llevar a cabo desde un recurso de tipo Dataset.

Funciones Básicas

- Resumen estadístico.
- Selección de atributos.
- Selección del atributo objetivo a clasificar (en el caso de clasificaciones).
- Creación de datasets a partir de otros datasets:
    - Muestreo
    - Separación en subconjuntos de entrenamiento y prueba.

Funciones Avanzadas

- Visualización interactiva a través de Dynamic Scatterplot
- Detección de anomalías.
- Creación de datasets a partir del filtrado.
- Creación de nuevos atributos.

Aprendizaje Supervisado vs No supervisado

El vídeo con este nombre de la serie de vídeos educacionales de BigML contiene un resumen muy breve sobre las características fundamentales de ambos tipos de aprendizaje. A lo largo de la asignatura hemos tratado este tema con mucha mayor profundidad y por eso no vemos necesario incluir un resumen de dicho vídeo en este documento.

Profundización: OptiML

Conocer qué algoritmo es adecuado aplicar para cada dataset es uno de los desafíos del Machine Learning, pero no es el único, también hay que afinar los hiperparámetros del algoritmo para obtener el resultado óptimo.

Precisamente, el objetivo de OptiML es facilitar ambas tareas:

1. Elección del algoritmo supervisado más adecuado al dataset.
2. Afinación de los hiperparámetros del algoritmo (hand-tuning),

Para ello, OptiML crea y evalúa múltiples modelos con diferentes configuraciones, sirviéndose de la optimización de parámetros bayesiana.

Configuración

A continuación, proporcionamos un listado con los aspectos a configurar por el usuario a la hora de utilizar OptiML.

Configuración básica:

- Atributo objetivo de la clasificación
- Máximo tiempo de entrenamiento.
- Número de modelos a evaluar.

Configuración avanzada:

- Tipos de modelos a incluir en la evaluación.
- Método de evaluación.
- Métrica para optimizar
- Pesos
- Valores nulos
- Muestreo

Fases de OptiML

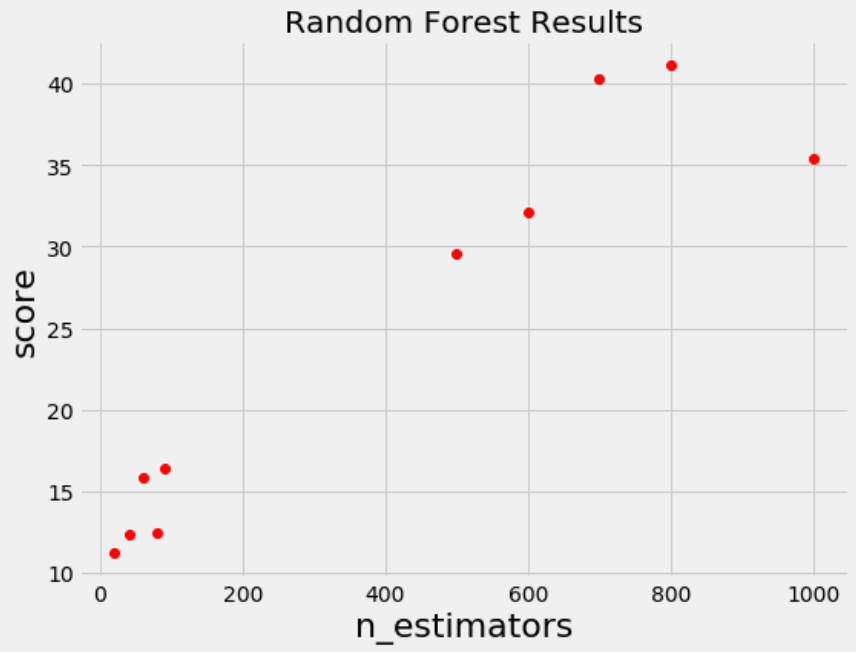
1. **Búsqueda de parámetros:**

Método Hold-out. Se selecciona aleatoriamente una parte de los datos para llevar a cabo la búsqueda.

Se sirve de la optimización bayesiana de parámetros. Este tipo de métodos son muy eficientes ya que en la búsqueda iterativa de hiperparámetros óptimos, cada iteración se basa, a diferencia de los métodos manuales o aleatorios, en los resultados de las iteraciones anteriores.

1. Creamos un modelo probabilístico sustituto de la función objetivo.
2. Encontramos los hiperparámetros óptimos para el modelo sustituto.
3. Aplicamos los hiperparámetros al modelo real objetivo.
4. Actualizamos el modelo probabilístico con los resultados del modelo real.
5. Repetimos los pasos 1.2, 1.3 y 1.4 hasta llegar al tiempo/iteraciones máximas.

A modo de ejemplo, en la siguiente gráfica (Will Koehrsen, 2018), si el modelo es más preciso cuanto menor sea el eje de coordenadas *score*, un modelo de optimización bayesiana de parámetros seguirá buscando el hiperparámetro n_estimators óptimo cerca de aquellos que han tenido mayor éxito previamente (es decir, en el intervalo entre 0 y 200).

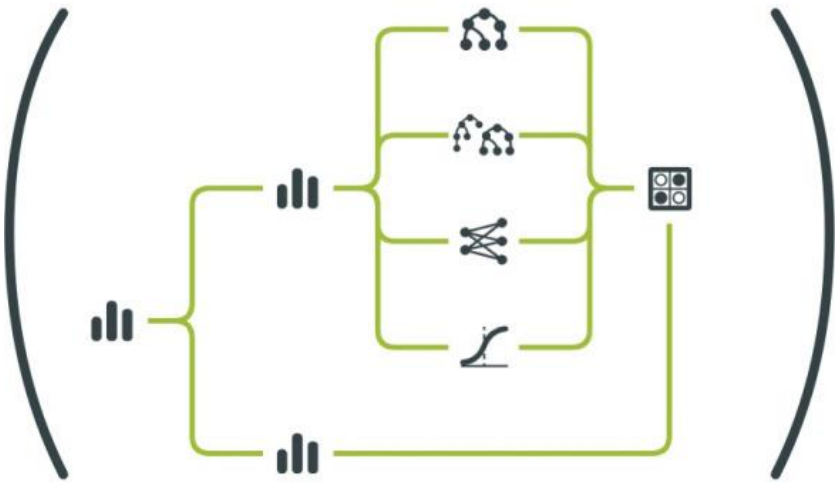


Lower Score being better [Will Koehrsen, 2018]

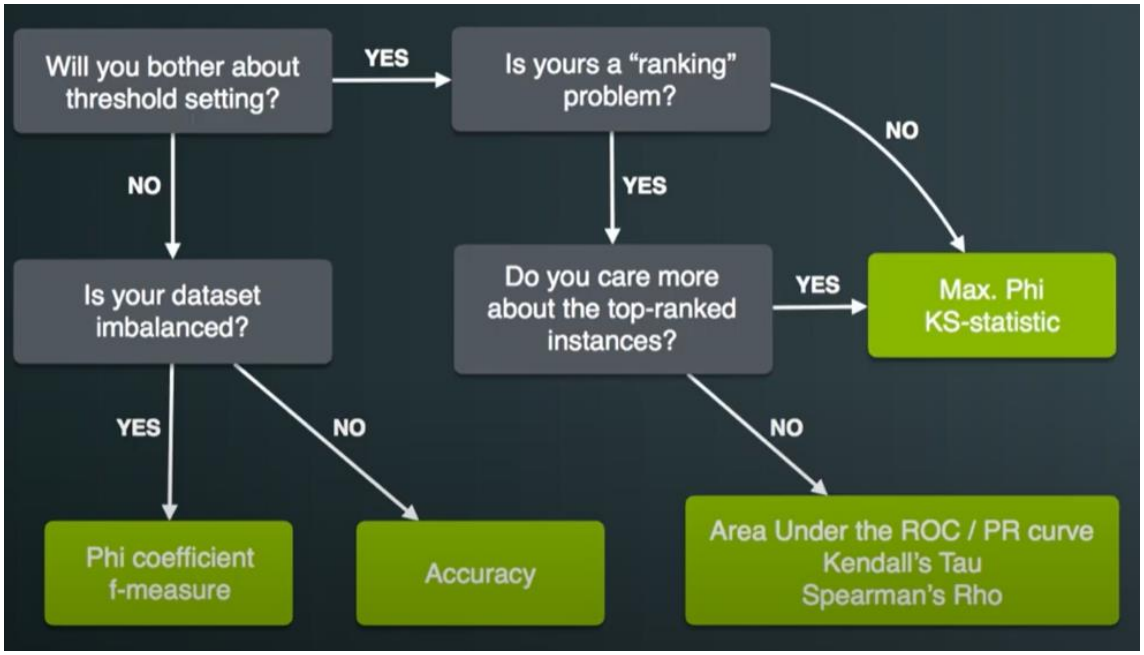
2. Validación

Cuando se hayan encontrado los hiperparámetros más prometedores para todos los modelos contemplados, se lleva a cabo una fase de validación con el dataset completo. Para esta validación se utiliza el método de validación cruzada, que no se ha utilizado en la fase uno puesto que es un método costoso desde el punto de vista computacional.

.



Elección de métrica del rendimiento del modelo



Selección del modelo

El resultado final de OptiML es un conjunto de modelos, con sus correspondientes hiperparámetros y ordenados según la métrica definida para el rendimiento.

A la hora de seleccionar el método, BigML nos alerta de que según la naturaleza del problema que queramos resolver, no siempre debemos escoger el modelo que haya tenido mejor rendimiento.

Debemos tener en cuenta otras características a nivel de complejidad del modelo que pueden quedar resumidas, a grandes rasgos, en la siguiente tabla:

Métodos Sencillos	Métodos Complejos
Débiles	Lentos
Sesgados	Necesitan gran cantidad de datos
Interpretabilidad	Representabilidad
Confianza	Rendimiento

Otras preguntas que deberíamos hacernos a la hora de seleccionar el modelo

- ¿Cambiará la cantidad de datos a lo largo del tiempo y habrá que **reentrenar** el modelo?
- ¿Necesitamos que el modelo no varíe frente a pequeños cambios de datos?
- ¿Cuál es la latencia máxima aceptable?

Bibliografía

BigML Education Videos | BigML.com

https://www.youtube.com/watch?list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&v=6xbNpILmQYo&feature=youtu.be&ab_channel=bigmlcom

https://www.youtube.com/watch?list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&v=YC6LliGR2Io&feature=youtu.be&ab_channel=bigmlcom

https://www.youtube.com/watch?list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&v=1AVrWvRvfxs&feature=youtu.be&ab_channel=bigmlcom

https://www.youtube.com/watch?list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&v=1AVrWvRvfxs&feature=youtu.be&ab_channel=bigmlcom

https://www.youtube.com/watch?list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&v=rGIsJ9gJfac&feature=youtu.be&ab_channel=bigmlcom

BigML Blog

https://blog.bigml.com/2018/05/08/introduction-to-optiml-automatic-model-optimization/

Will Koehrsen, 2018. A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning. Towardsdatascience.com

https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f