# Predicción de tipos de suelos forestales

Juan A. Botía Blaya
juanbot@um.es

November 27, 2007

## 1 Introducción

Dentro de los ingenieros agrónomos, podemos encontrar a los gestores de recursos naturales. Estos son responsables del desarrollo de estragias de gestión para el desarrollo de ecosistemas. Para la creación de estas estrategias de gestión, son necesarios varios elementos, de entre los que podemos destacar datos de inventario sobre suelos de bosque. Estos datos son un elemento crucial para su proceso de decisión. En este contexto, existe una dificultad añadida derivada del hecho de que un ingeniero forestal está a cargo de una cierta zona de bosque, lo cual implica que la disponibilidad de datos se circunscribe a esa zona. Sin embargo, el disponer de datos sobre el tipo de cubierta forestal de los suelos adyacentes a su zona puede aumentar la capacidad del proceso de decisión. Pues bien, un método para obtener esta información (o al menos estimarla), es generando modelos predictivos.

El objetivo de esta práctica es la generación de modelos predictivos de dichas cubiertas en suelo forestal, haciendo uso de, al menos 3 de los algoritmos que incorpora la herramienta Weka. Una vez generados esos modelos, se deberá presentar una comparativa de los mismos, con los métodos de estimación de error que hemos visto en clase (e.g. hold-out, validación cruzada o bootstrapping).

El área que se ha estudiado comprende cuatro regiones salvajes que se localizan en el Parque Nacional Roosevelt, en el estado de Colorado, EEUU. Cada una de las muestras representa una celda de $30 \times 30$ celdas. El tipo de suelo para cada una de las celdas se generó a partir del uso del sistema RIS (*Resource Information System*) del sistema forestal de los EEUU (USFS). En el correspondiente conjunto de datos, se pueden encontrar 12 variables independientes. El valor para esas variables se obtuvo a partir del US Geological Survey (USGS) y también del USFS.

Las cuatro regiones se denominan *Neota*, *Rawah*, *Comanche Peak* y *Cache la Poudre*. Con respecto a los niveles de altura a los que se encuentran cada una, la primera es la más alta (i.e. su nivel medio de elevación es el más alto de los cuatro). La segunda y la tercera, ambas están aproximadamente al mismo nivel y por debajo de la primera. La cuarta es la más baja.

Si hablamos de los tipos de cubierta vegetal que va a tener cada una de las cuatro zonas, `Neota` tendría del tipo 1, la segunda y la tercera tendría sobre todo del tipo 2, seguidas del tipo 1 y 5. La cuarta tendría cubierta de los tipos 3, 6 y 4.

La descripción de todos los atributos del conjunto de datos aparece en el fichero `memoria.info`.

## 2 Metodología

Para la resolución de la práctica, se hará necesario un estudio basado en estadística descriptiva de los atributos. Tanto de las características de entrada como de salida. De esta forma, previo al planteamiento de hacer uso de cualquier algoritmo, se deberá generar un informe similar al que usamos como memoria para la primera práctica.

Una vez hecho esto, haremos uso de ese informe para respaldar decisiones que podamos tomar a cerca de la limpieza de los datos como por ejemplo, no usar todas las muestras o eliminar atributos que no aporten capacidad discriminante.

Se han de elegir tres algoritmos de clasificación, justificar su elección con una argumentación que justifique el que sean idoneos para este problema concreto. Una vez hecho esto, se ha de generar un informe de la aplicación de cada uno de los algoritmo por separado que incluya:

- decisiones de configuración de parámetros del algoritmo (e.g. el valor de un ratio de aprendizaje $\alpha$ o el número de nodos ocultos de una red neuronal),
- condiciones de ejecución del mismo,
- Si se han hecho diferentes experimentos, indicarlos explicando cada uno de ellos,
- Medidas de error, matrices de confusión, etc.

Y posteriormente, un informe conjunto que explique cuál de ellos, si es que eso es así, es mejor que el resto y por qué. Por tanto, el material a entregar al profesor se reduce a un documento que incluya el análisis exploratorio de los datos, el estudio de cada algoritmo y el estudio de rendimiento conjunto. La fecha límite de entrega es el 10 de enero. TODAS LAS ENTREGAS HAN DE SER POR CORREO ELECTRÓNICO. El formato de la memoria será `pdf`. Las defensas de los trabajos serán realizadas el 17 de enero en horario lectivo.