UNIVERSITAS Miguel Hernández

Tema 6. Random Forest.

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones.

Índice

Descripción de la técnica

- Se realizan múltiples árboles de decisión o de regresión (de ahí lo de 'Forest').
- Para la predicción o clasificación, el individuo pasa por todos los árboles.
- En clasificación: se clasifica en la categoría con más votos.
- En regresión: se predice la media de los resultados.

Construcción de los Árboles

- Se seleccionan aleatoriamente N individuos de la muestra de entrenamiento, con reemplazamiento (bootstrap).
- Para cada árbol, algunos individuos se repiten y otros no son seleccionados.

Selección de Variables

- Se especifica un número m de variables entre M variables.
- Para cada árbol, se seleccionan m variables aleatorias.
- Cada árbol crece hasta su máxima extensión sin poda, incluso si hay nodos con pocos individuos.

Parámetros

- número de árboles en el bosque.
- número de variables aleatorias m.
- mínimo número de registros por nodo.
- número máximo de nodos terminales u hojas.

Ventajas

- Pocas suposiciones.
- Mínima preparación de datos (convertir variables categóricas a numéricas y viceversa).
- Maneja miles de variables y selecciona las más importantes sin reducir la dimensionalidad.
- Se puede usar como método no supervisado (clustering, detección de outliers).

Desventajas

- Pérdida de interpretación: el modelo es una 'caja negra' (similar a las redes neuronales).
- No se comporta bien para predecir fuera del rango de los datos de entrenamiento.


UNIVERSITAS
Miguel Hernández