Aprendizaje Automático

Departamento de Informática – UC3M

TUTORIAL 6 – Random Forest

2025

Tutorial 6

Recordando teoría. Random Forest

- Ejemplo

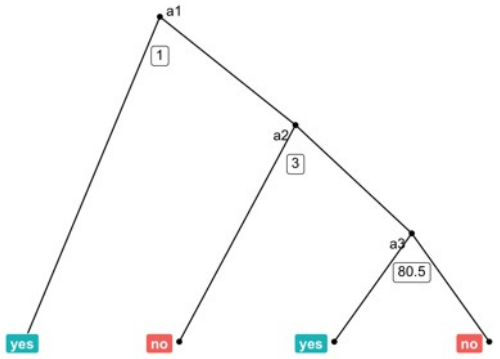
#	a_1	a_2	a_3	a_4	Jugar
1	overcast	hot	86	FALSE	yes
2	overcast	cool	65	TRUE	yes
3	overcast	mild	90	TRUE	yes
4	overcast	hot	75	FALSE	yes
5	rainy	mild	96	FALSE	yes
6	rainy	cool	80	FALSE	yes
7	rainy	cool	70	TRUE	no
8	rainy	mild	80	FALSE	yes
9	rainy	mild	91	TRUE	no
10	sunny	hot	85	FALSE	no
11	sunny	hot	90	TRUE	no
12	sunny	mild	95	FALSE	no
13	sunny	cool	70	FALSE	yes
14	sunny	mild	70	TRUE	yes

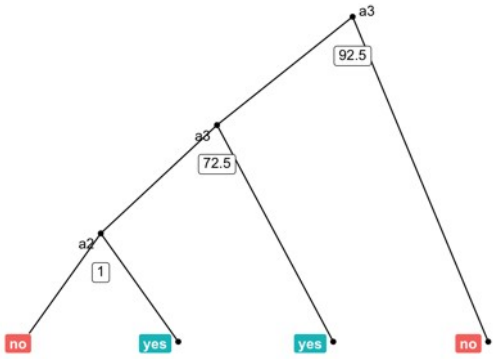
- Para 4 árboles
  - Muestreo:

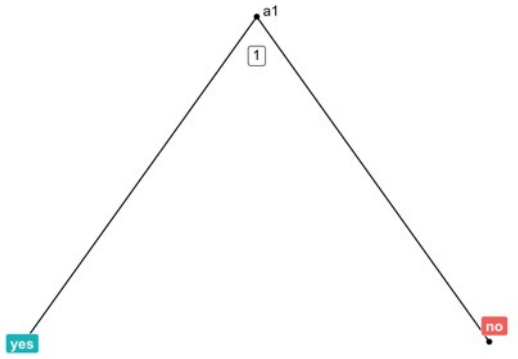
Instancias de test para evaluar el error (*out of bag*)

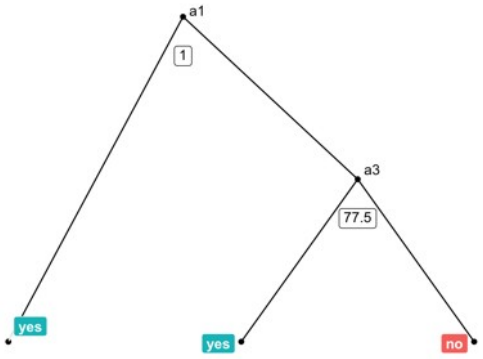
	Instancias Seleccionadas														Atributos Seleccionados	OOB
L_1	6	8	9	13	6	7	12	14	7	3	5	4	5	13	a_1, a_2, a_3	1,2,10,11
L_2	9	5	10	3	11	1	7	14	9	11	5	3	14	3	a_2, a_3	2,4,6,8,12,13
L_3	6	1	12	2	10	13	4	14	11	9	1	7	3	5	a_1	8
L_4	8	4	12	14	2	8	13	1	5	4	7	13	8	3	a_1, a_3	6,9,10,11

  - Creación:









Tutorial 6

Random Forest. Regresión/Clasificación

pros

- Puede manejar instancias con cientos de atributos (numéricos/categóricos)
- Se reduce el sesgo y la varianza
- Requieren menos limpieza (no requieren estandarización)
- Se obtiene la importancia de los atributos (pureza, permutación)
- No se ven influenciados por outliers
- Estimación de probabilidades: puede cuantificar el grado de creencia en una predicción: porcentaje de árboles que dan una respuesta
- OOB estimar el error durante el aprendizaje, sin CV

con

- Sesgo hacia atributos categóricos con muchos valores frente a pocos
- Dificultad para explicar las predicciones
- No son capaces de extrapolar fuera del rango de los predictores observado en los datos de entrenamiento

Tutorial 6

Random Forest. Parámetros sklearn

- **n_estimators**: número de árboles incluidos en el modelo.
- **max_features**: número de predictores considerados a en cada división. Puede ser:
  - Un valor entero
  - Una fracción del total de predictores. Se calcula como max(1, int(max_features * n_features_in_)). Si su valor es 1.0 tiene en cuenta todos los predictores.
  - “sqrt”, raiz cuadrada del número total de predictores.
  - “log2”, log2 del número total de predictores.
  - None, utiliza todos los predictores (igual que 1.0)
- **oob_score**: Si se calcula o no el out-of-bag R^2. Por defecto es False ya que aumenta el tiempo de entrenamiento.
- Otros: max_depth, min_samples_split, min_samples_leaf, ...

Tutorial 6

Extremely Randomized Trees (Extra-Trees)

- Extra-Trees son un paso más en la aleatoriedad
  - Igual que RF toman un subconjunto de atributos aleatorios
  - A diferencia, se genera un valor de corte aleatorio para cada atributo y se toma el mejor
  - El muestreo de instancias es sin reemplazo (*boostrap*)

Random Forest	**Extra-Trees**
Selección de subconjunto de instancias	Selección de instancias del conjunto
División por un valor optimizado (Gini)	División por un valor aleatorio
Varianza media	Varianza baja
Costoso encontrar el mejor nodo para hacer la división	Rápido para hacer divisiones

Tutorial 6

Extremely Randomized Trees (Extra-Trees). Parámetros sklearn

- Los mismos que Random Forest, añadimos dos de interés:
  - criterion ={“squared_error”, “absolute_error”, “friedman_mse”, “poisson”}, default=”squared_error” Mide la calidad de la partición realizada.
  - bootstrapbool, default=False. Si se utiliza bootstraping en la selección de muestras. Con el valor de False, se utilizan todas las muestras.