Jesús García Herrero

TÉCNICAS DE REGRESIÓN NO LINEAL

En esta clase se presenta un método de inducción de modelos numéricos de regresión a partir de datos. En el tema de técnicas clásicas se presentó la regresión lineal como técnica muy potente para determinar funciones lineales de manera óptima y con un procedimiento eficiente y escalable. Sin embargo, la limitación clara es la linealidad, aspecto que se ha intentado superar con diferentes aproximaciones.

Se presentan aquí las estrategias de inducción de árboles de predicción numérica, distinguiéndose los árboles de regresión, que aproximan la función mediante tramos constantes, y los árboles de modelos, que aproximan con una serie de modelos lineales por tramos. Aquí se presenta el método M5, detallando las fases que lleva a cabo para construir el modelo numérico:

Construcción del árbol con heurísticas para determinar cada atributo que mejor divide los datos en un proceso recursivo

- Construcción de modelos lineales en las hojas y nodos del árbol
- Poda de los modelos lineales, eliminando atributos no significativos
- Poda del árbol, eliminando niveles con el objetivo de mejorar la generalización y continuidad del modelo global

El proceso constructivo se ilustra con ejemplos y para concluir se presentan aspectos prácticos que aparecen en conjuntos de datos reales: datos incompletos y mezcla de datos numéricos y datos nominales.

.

Predicción numérica

Técnicas de regresión y predicción de datos

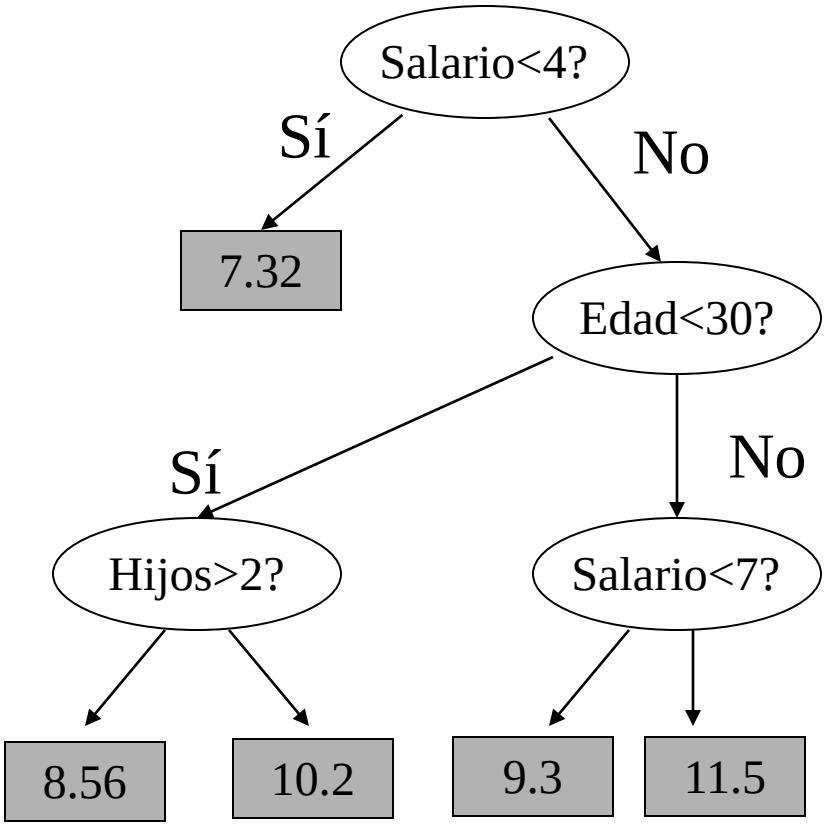
Jesús García Herrero
Universidad Carlos III de Madrid

Universidad
Carlos III de Madrid

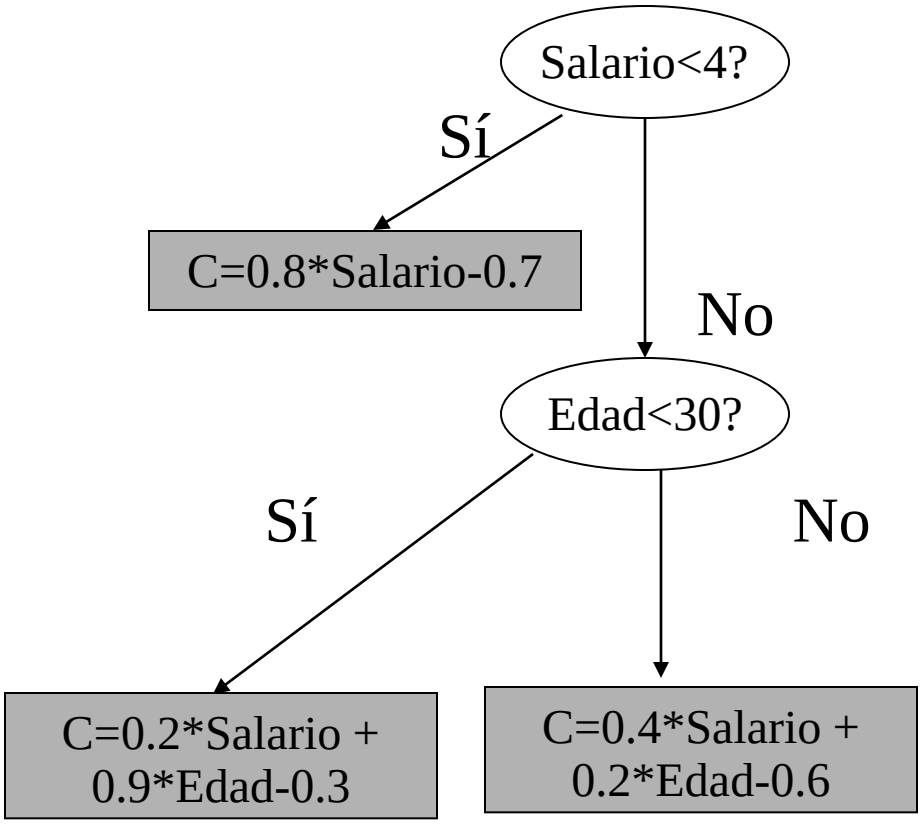
Árboles para predicción numérica

- Muchas veces las clases son contnuas (bolsa, CPU,…)
- Clásicamente, se ha utilizado la **regresión lineal**, pero
  - los modelos obtenidos sólo operan con atributos numéricos
  - se impone una dependencia puramente lineal
- Los **árboles de regresión** se generan de forma similar a los de decisión, con valores promedio en cada hoja. CART (Breiman 84)
- M5 (Quinlan, 93) es una variación de CART que genera modelos lineales en las hojas, **árbol de modelos**, en lugar de valores numéricos.
- El nuevo heurístico para separar ejemplos no es la entropía de clases, sino la varianza del error en cada hoja

Árboles para predicción numérica



árbol de regresión



árbol de modelos

Árboles para predicción numérica

Construcción del árbol

- Heurística: minimizar la variación interna de los valores de la clase dentro de cada subconjunto
- Medida concreta: elegir aquel atributo que maximice la reducción de la desviación estándar del error (SDR), con la fórmula:

$$SDR = SD(E) - \sum_j \frac{|E_j|}{|E|} SD(E_j)$$

donde:

  - E es el conjunto de ejemplos en el nodo a dividir
  - Ej son los ejemplos con valor j del atributo considerado
  - |.| es el número de ejemplos en cada conjunto
  - SD(C) es la desviación típica de los valores de la clase en C
- Finalización: pocos ejemplos (2), o poca variación de los valores de clase (5% del valor en el conjunto de instancias original)

Árboles para predicción numérica

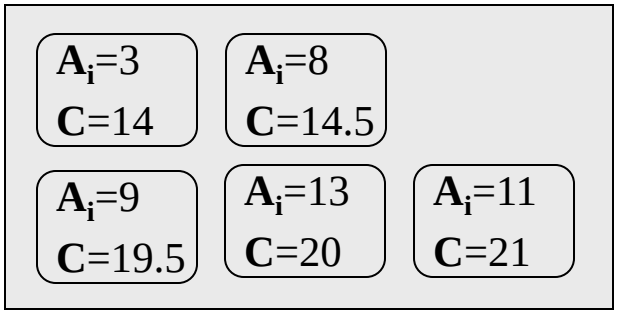
Estimación del error

- Cálculo de la desviación estándar del error en un conjunto I:

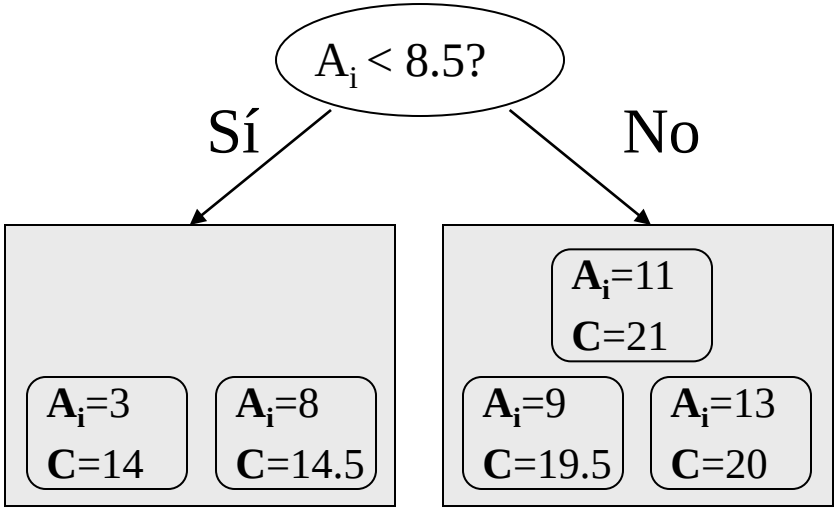
$$e(I) = \frac{1}{n} \sum_{i \in I} \left\| c(i) - c(m,i) \right\|$$

  donde:
  - n=|I|
  - c(i) es la clase de la instancia i
  - c(m,i) es la clase predicha con el modelo m para la instancia i
    - durante la construcción del árbol, el modelo es el valor medio en cada hoja
    - en la poda, se construye un modelo MC para predecir
- hay versiones con desviación estándar y otras con suma de módulos

Ejemplo



SD=2.94



$SD(A_i)=2/5*0.25+3/5*0.62=0.47$

Árboles para predicción numérica

Construcción del árbol

- Ejemplos con faltas

$$SDR = \frac{m}{|E|}\left[SD(E) - \sum_j \frac{|E_j|}{|E|} SD(E_j) \right]$$

  m es el número de instancias sin faltas en ese atributo
- Atributos nominales
  - Un atributo con k valores se transforma en k-1 atributos binarios:
    - se ordenan los valores según el valor de la clase
    - cada atributo binario es un posible punto de separación: 0,1
  - Ej: Motor = {A, B, C, D}, con clases 10, 11, 8, 7
    - D vs C, A, B
    - D, C vs A, B
    - D, C, A vs B
  - Todos las decisiones son binarias!

Árboles para predicción numérica

Poda del árbol

- Primero se obtiene un modelo de regresión lineal para cada uno de los nodos interiores del árbol no podado:

  $a_0+a_1x_1+a_2x_2+….+a_kx_k$

  – x_1, x_2,…, x_k son los atributos que aparecen únicamente en el subárbol que hay por debajo

  – para los atributos nominales, cada atributo sintético binario puede tomar valores {0,1}

- El subárbol se poda si el error sobre nuevas instancias es inferior con el modelo que con él

  – heurístico: (n+v)/(n-v)

    - n es el número de instancias que alcanzan el nodo
    - v es el número de parámetros en el modelo lineal

  – permite simplificar el modelo quitando atributos

Árboles para predicción numérica

Pseudocódigo M5'

```
M5'(ejemplos)
{   SD=sd(ejemplos)
    para cada k-val atributo nom.
            crear k-1 atributos binarios
    raiz=nuevo nodo
    raiz.ejemplos=ejemplos
    dividir(raiz)
    podar(raiz)  }

podar (nodo)
{   si (nodo.tipo=INTERIOR)
            podar (nodo.hijoIzquierdo)
            podar (nodo.hijoDerecho)
            nodo.modelo=regresLineal(nodo)
            si errorSubarbol(nodo)>error(nodo)
            nodo.tipo=HOJA  }
```

```
dividir (nodo)
{   si (size(nodo.ejemplos)<4 o
    SD(nodo.ejemplos)<0.05*SD)
            nodo.tipo=HOJA
    si no
            nodo.tipo=INTERIOR
            para cada atributo
                    para cada posible división
                            calcular SDR
            nodo.atributo=atributo con máximo
                            SDR
            dividir (nodo.izquierdo)
            dividir (nodo.derecho) }
```

Árboles para predicción numérica

Ejemplo

relation elusage (64 ejemplos)

- attribute average_temperature integer
- attribute month { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
- attribute average_electricity_usage real

73,8,24.828
67,9,24.688
57,10,19.31
43,11,59.706
26,12,99.667
41,1,49.333
38,2,59.375
…

Árboles para predicción numérica

```
average_temperature <= 54.5 :
|   average_temperature <= 38.5 :
|   |   average_temperature <= 33 :
|   |   |   average_temperature <= 29 : LM1 (2/4.12%)
|   |   |   average_temperature >  29 : LM2 (2/77.1%)
|   |   average_temperature >  33 :
|   |   |   month=12,1 <= 0.5 : LM3 (2/20.9%)
|   |   |   month=12,1 >  0.5 :
|   |   |   |   month=1 <= 0.5 : LM4 (3/60.5%)
|   |   |   |   month=1 >  0.5 : LM5 (3/52.6%)
|   average_temperature >  38.5 :
|   |   average_temperature <= 44 :
|   |   |   month=12,1 <= 0.5 :
|   |   |   |   average_temperature <= 40 : LM6 (2/53.7%)
|   |   |   |   average_temperature >  40 : LM7 (3/43.7%)
|   |   |   month=12,1 >  0.5 : LM8 (2/74.9%)
|   |   average_temperature >  44 :
|   |   |   month=4,11,3,2,12,1 <= 0.5 : LM9 (2/6.7%)
|   |   |   month=4,11,3,2,12,1 >  0.5 :
|   |   |   |   average_temperature <= 50 :
|   |   |   |   |   average_temperature <= 46.5 : LM10 (2/18%)
|   |   |   |   |   average_temperature >  46.5 : LM11 (3/23.1%)
|   |   |   |   average_temperature >  50 : LM12 (2/1.71%)
```

```
average_temperature >  54.5 :
|   month=9,8,5,10,4,11,3,2,12,1 <= 0.5 :
|   |   average_temperature <= 76 :
|   |   |   month=7,9,8,5,10,4,11,3,2,12,1 <= 0.5 :
|   |   |   |   average_temperature <= 71.5 : LM13 (2/6.97%)
|   |   |   |   average_temperature >  71.5 : LM14 (2/17.2%)
|   |   |   month=7,9,8,5,10,4,11,3,2,12,1 >  0.5 : LM15 (2/13.3%)
|   |   average_temperature >  76 : LM16 (2/2.21%)
|   month=9,8,5,10,4,11,3,2,12,1 >  0.5 :
|   |   average_temperature <= 72 :
|   |   |   month=10,4,11,3,2,12,1 <= 0.5 :
|   |   |   |   month=8,5,10,4,11,3,2,12,1 <= 0.5 : LM17 (5/18.8%)
|   |   |   |   month=8,5,10,4,11,3,2,12,1 >  0.5 :
|   |   |   |   |   average_temperature <= 62 : LM18 (2/22%)
|   |   |   |   |   average_temperature >  62 : LM19 (3/19.4%)
|   |   |   month=10,4,11,3,2,12,1 >  0.5 :
|   |   |   |   average_temperature <= 56.5 : LM20 (3/19.6%)
|   |   |   |   average_temperature >  56.5 : LM21 (2/8.71%)
|   |   average_temperature >  72 : LM22 (4/5.38%)
```

Models at the leaves:

LM1: average_electricity_usage = 100
LM2: average_electricity_usage = 86.4
LM3: average_electricity_usage = 63.5
LM4: average_electricity_usage = 66.4
LM5: average_electricity_usage = 76.4
LM6: average_electricity_usage = 55.2
LM7: average_electricity_usage = 48.6
LM8: average_electricity_usage = 64.2
LM9: average_electricity_usage = 42.3
LM10: average_electricity_usage = 51.6
LM11: average_electricity_usage = 44.4
LM12: average_electricity_usage = 55.2
LM13: average_electricity_usage = 22.1
LM14: average_electricity_usage = 13.8
LM15: average_electricity_usage = 27
LM16: average_electricity_usage = 17.4
LM17: average_electricity_usage = 24.4
LM18: average_electricity_usage = 30.3
LM19: average_electricity_usage = 27.1
LM20: average_electricity_usage = 24.8
LM21: average_electricity_usage = 21
LM22: average_electricity_usage = 23.5

Ejemplo. Árbol de modelos

Pruned training model tree:

average_temperature <= 54.5 :
| average_temperature <= 38.5 : LM1 (12/58.6%)
| average_temperature > 38.5 : LM2 (16/45.6%)
average_temperature > 54.5 :
| month=9,8,5,10,4,11,3,2,12,1 <= 0.5 : LM3 (8/15%)
| month=9,8,5,10,4,11,3,2,12,1 > 0.5 : LM4 (19/20.7%)

Models at the leaves:

LM1: average_electricity_usage = 169 - 2.74average_temperature
LM2: average_electricity_usage = 49.1 + 15.1month=12,1
LM3: average_electricity_usage = 181 - 2.28average_temperature
+ 14.5month=7,9,8,5,10,4,11,3,2,12,1
LM4: average_electricity_usage = 25

Árboles para predicción numérica

Ejemplo. Árbol de regresión

Pruned training regression tree:

average_temperature <= 54.5 : LM1 (28/94.2%)
average_temperature > 54.5 : LM2 (27/25.5%)

Models at the leaves:

Unsmoothed (simple):

LM1: average_electricity_usage = 62.3
LM2: average_electricity_usage = 23.5

Árboles para predicción numérica

Ejemplo

Árbol de regresión

=== Evaluation ===

=== Summary ===

Correlation coefficient	0.8155
Mean absolute error	9.6511
Root mean squared error	13.7671
Relative absolute error	48.7555 %
Root relative squared error %	57.8696
Total Number of Instances	55

Árbol de modelos

=== Evaluation ===

=== Summary ===

Correlation coefficient	0.9454
Mean absolute error	5.7019
Root mean squared error	7.7566
Relative absolute error	28.8051 %
Root relative squared error %	32.6045
Total Number of Instances	55

Árboles para predicción numérica

Suavizado para predicción

- Suele ser beneficioso evitar las discontinuidades mediante un proceso de suavizado, sobre todo con pocos ejemplos
  - Se construyen modelos para todos los nodos (hojas e internos)
  - El valor en el nodo hoja se filtra hacia atrás hasta el nodo raíz
- Predicción del valor de un ejemplo de test:
  - se sigue el árbol hasta el modelo en la hoja correspondiente, y se obtiene el valor
  - A continuación, se propaga hacia arriba el valor hasta la raíz:

$$p' = \frac{np + kq}{n + k}$$

  - p: predicción que llega a este nivel
  - p': predicción filtrada, para enviar hacia arriba
  - n: número de ejemplos que alcanzan el nodo inferior
  - k: factor de suavizado