Jorge Sossa
Gary Vidal

Proyecto Final : Árboles de clasificación/regresión Bayesiana.

20 de Octubre de 2022

1. Descripción

Este proyecto se centra en la creación de modelos de clasificación, ajustando de manera estocástica un árbol de decisión.

El problema a resolver es encontrar un árbol de decisión que se ajuste al modelo, de tal manera que las predicciones sean lo más precisas posible.

2. Método

A diferencia de otros maneras para ajustar árboles, este método utiliza dos ejes principales:

- Se especifica una distribución a priori.
- Se realiza una búsqueda estocástica para encontrar el posteriori.

La distribución a priori de un árbol $\mathbb{P}(T)$ se obtiene identificando los nodos terminales y los nodos interiores, a los cuales se les asigna una probabilidad dependiendo de su altura y si estos fueron divididos o no, la probabilidad $\mathbb{P}(T)$ es la multiplicación de todos estos términos.

2.1. Algoritmo

La búsqueda del posteriori se realiza con el algoritmo de Metropolis-Hasting, donde se busca crear una cadena $T_0, T_1, ...$ de árboles que converge hacia la distribución a posteriorí.

Definiendo los conjuntos :

· X : Conjunto de predictores (características de los datos)

· Y : Clases de los datos.

Algorithm 1 Algoritmo M-H

1: Comenzando del árbol trivial T^0 (solo un nodo).
2: Generamos un candidato T^* con probabilidad $q(T^i, T^*)$
3: Calculamos $u \sim U[0,1]$ y $\alpha(T^i, T^*) = min\left\{\frac{q(T^*,T^i)\mathbb{P}(Y|X,T^*)\mathbb{P}(T^*)}{q(T^i,T^*)\mathbb{P}(Y|X,T^i)\mathbb{P}(T^i)}, 1\right\}$
4: **if** $u \leq \alpha$ **then**
5: tomamos $T^{i+1} = T^*$ y volvemos al paso 3)
6: **else**
7: $T^{i+1} = T^i$ (La cadena no cambia)
8: **end if**

3. Resultados

Con el dataset "Breast cancer" de Sklearn, el algoritmo obtuvo buenos resultados en pocas iteraciones, teniendo una precisión de sobre el 90 % vs un 85 % de precisión del método con Decision Trees de Sklearn. :

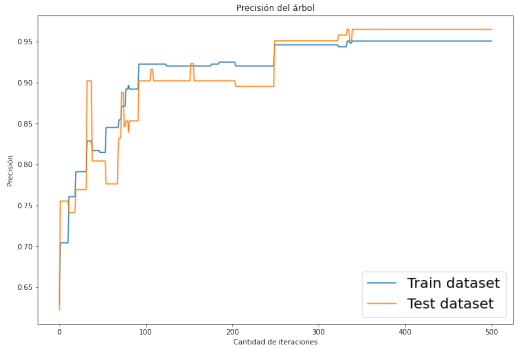

Figura 1: Precisión del algoritmo luego de 500 iteraciones.

4. Bibliografía :

- Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. Journal of the American Statistical Association, 93 (443), 935-948. http://dx.doi.org/10.1080/01621459.1998.10473750
- A Practical Markov Chain Monte Carlo Approach to Decision Problems. Timothy Huang, Yuriy Nevmyvaka