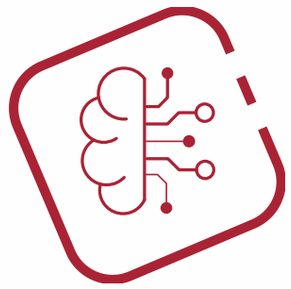
ITSAR ISTITUTO TECNICO SUPERIORE
ANGELO RIZZOLI

Algoritmi per il Machine Learning

Ing Andrea Colleoni



Esempio

	Gender	Height	Weight	Index
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3

As in all algorithms, the cost function is the basis of the algorithm. In the case of decision trees, there are two main cost functions: the Gini index and entropy.
Any of the cost functions we can use are based on measuring impurity. Impurity refers to the fact that, when we make a cut, how likely is it that the target variable will be classified incorrectly.

In the example above, impurity will include the percentage of people that weight >=100 kg that are not obese and the percentage of people with weight<100 kg that are obese. Every time we make a **split** and the classification is not perfect, the split is impure.

Information Gain Criterion

Dobbiamo dividere il dataset in modo che l'entropia attesa dei dataset risultanti sia la minima possibile.

- Il criterio di minimizzazione dell'entropia attesa e detto «Information Gain Criterion». Tecnicamente, l'information gain è la diminuzione dell'entropia passando dal dataset intero ai due sotto-dataset.
- Minimizzare l'entropia attesa equivale a massimizzare il guadagno informativo.
- Allo stesso modo, potremmo utilizzare l'impurità di Gini GI come criterio da minimizzare.

Impurità attesa $$E_{j,\theta}(\mathrm{H}) = \frac{|D_{x_j<\theta}|}{|D|}\mathrm{H}(Y|x_j < \theta) + \frac{|D_{x_j\geq\theta}|}{|D|}\mathrm{H}(Y|x_j \geq \theta).$$

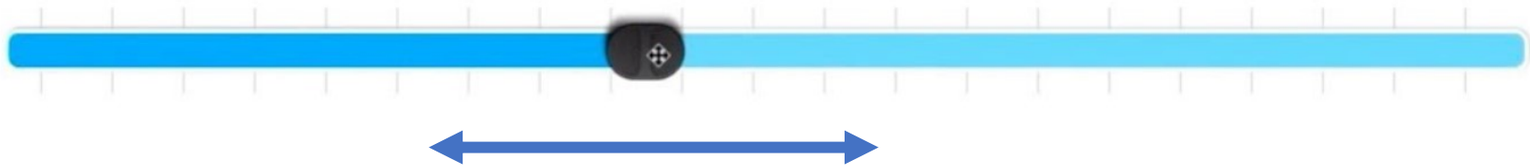
Information Gain $$IG(T,a) = \mathrm{H}(T) - \mathrm{H}(T|a)$$

IG: esempio

Outlook	Temperature Numeric	Temperature Nominal	Humidity Numeric	Humidity Nominal	Windy	Play
overcast	83	hot	86	high	FALSE	yes
overcast	64	cool	65	normal	TRUE	yes
overcast	72	mild	90	high	TRUE	yes
overcast	81	hot	75	normal	FALSE	yes
rainy	70	mild	96	high	FALSE	yes
rainy	68	cool	80	normal	FALSE	yes
rainy	65	cool	70	normal	TRUE	no
rainy	75	mild	80	normal	FALSE	yes
rainy	71	mild	91	high	TRUE	no
sunny	85	hot	85	high	FALSE	no
sunny	80	hot	90	high	TRUE	no
sunny	72	mild	95	high	FALSE	no
sunny	69	cool	70	normal	FALSE	yes
sunny	75	mild	70	normal	TRUE	yes

Partizione ricorsiva del dataset

- Per ottenere un albero, detto «albero di decisione», possiamo applicare ripetutamente il passo induttivo a ciascun sotto-dataset, riducendo di volta in volta l'impurità.
- La procedura termina «naturalmente» quando il sotto-dataset associato a un nodo contiene un solo elemento: a questo punto la variabile casuale associata all'output ha un solo valore quindi e necessariamente pura.



Condizione di terminazione

Per limitare la profondità dell'albero risultante, e possibile introdurre ulteriori criteri di terminazione per il passo induttivo:

- quando la profondità dell'albero raggiunge un limite massimo;
- quando l'impurità di un nodo e inferiore a una soglia predefinita;
- quando il numero di elementi nel sotto-dataset associato a un nodo e inferiore a un minimo predefinito.

Algoritmo

Dette:

- x = {x1, …, xj}: il vettore di attributi
- Θ : valore soglia di un attributo xj in cui partizionare il set di dati
- y il vettore obiettivo

1. Se il dataset D soddisfa un criterio di terminazione, non fare nulla. Altrimenti:
   a) per ogni combinazione j; Θ :
      I. Calcola i due sotto-dataset Dxj< Θ e Dxj >= Θ;
      II. Calcola l'impurità attesa;
      III. Se l'impurità è la minore trovata finora (IG più elevato), ricorda i valori ottimali j* e Θ*;
   b) Associa i parametri migliori (j* e Θ*) alla radice e genera due figli, sinistro e destro, associati rispettivamente a Dxj< Θ e Dxj>= Θ.
   c) Applica ricorsivamente la procedura ai figli sinistro e destro.