

Data & Web Mining

Claudio Lucchese claudio.lucchese@unive.it

K-NN Classifier

- **Cosa succede per k che tende a infinito?**
 - a. La predizione non è definita
 - b. L'algoritmo non termina
 - c. La predizione è la classe maggioritaria del dataset
- c: k tendente a infinito implica selezionare tutte le istanze del dataset per la predizione

K-NN Classifier

- **Cosa fa il classificatore a tempo di training?**
 - a. Costruisce un albero di decisione
 - b. Niente
 - c. Memorizza il Training Set (possibilmente indicizzandolo)
- c: per questo motivo viene classificato come Lazy Learner

K-NN Classifier

- **Che impatto ha la “scala” (min/max) delle features?**
 - a. Se le features sono su scala diversa la predizione è sicuramente sbagliata
 - b. Il calcolo della distanza è dominato dalla features con range [min-max] maggiore
 - c. Dipende dalla misura di distanza
 - d. Nessun impatto
- b, ma possiamo considerare come vera anche c perchè la misura di distanza potrebbe essere insensibile alla scala

Binary Decision Tree

- **Che impatto ha la “scala” (min/max) delle features?**
 - a. Se le features sono su scala diversa la predizione è sicuramente sbagliata
 - b. Il calcolo della distanza è dominato dalla features con range [min-max] maggiore
 - c. Dipende dalla misura di distanza
 - d. Nessun impatto
- d. il miglior partizionamento D_L e D_R non dipende dalla scala

Binary Decision Tree

- **Perchè “binary”?**

- a. Perché possono gestire solo problemi di classificazione binaria: vero/falso, giallo/blu, cerchio/quadrato
- b. Perché ogni nodo può avere 2 figli
- c. Perché ogni nodo può analizzare solo due features

- b.

Binary Decision Tree

- **Quante features possono essere coinvolte nel predicato di un nodo?**
 - a. Solo 1 features
 - b. Solo 2 features
 - c. Solo 1 se numerico, più di una se categorico
 - d. Dipende da _____
- d: dipende dalla nostra implementazione

Binary Decision Tree

- **Qual è la profondità massima?**
 - a. Non si può dire, perchè ?
 - b. Pari al numero di features
 - c. Pari al numero di istanze
 - d. Pari al prodotto tra numero di features e numero di istanze
- c: perché ciascuna foglia deve avere almeno un'istanza

Binary Decision Tree

- **Nel caso di variabile categorica con N valori distinti possibili, quali affermazioni sono vere?**
 - a. Ci sono N modi diversi di creare uno Split binario
 - b. Ci sono 2^N modi diversi di creare uno Split binario
 - c. Ci sono 2^{N-1} modi diversi di creare uno Split binario
 - d. Ci sono $2^{N-1}-1$ modi diversi di creare uno Split binario
- d: perchè i sottoinsiemi di N valori distinti sono 2^N , ma metà di questi sono complementare ed equivalenti (se la feature assume i valori ABCD, $x \in \{A\}$ è equivalente a $x \in \{BCD\}$), e infine lo split $x \in \{ \}$ non è valido.