

# Algoritmi Avanzati

Emanuele Nardi

24 agosto 2018

## Risposte alle domande a risposta multipla

1. Un sistema di machine learning si dice *overfitting* quando è troppo specializzato sui dati di addestramento e non è in grado di generare previsioni adeguate ai dati nuovi;
  2. In una regressione polinomiale di secondo grado in una variabile vanno determinati *tre coefficienti*;
  3. L'accuratezza indica il *numero di previsioni corrette* per un modello di classificazione;
  4. Il formato CSV è di tipo *strutturato*;
  5. È preferibile che gli insiemi di addestramento e di validazione *siano disgiunti* perché siamo interessati a valutare le prestazioni del sistema su esempi non visti durante l'addestramento;
  6. La *funzione sigmoide* mappa il valore reale in uscita dal regressore sull'intervallo  $[0, 1]$ ;
  7. La *funzione sigmoide* può essere definita come  $\frac{1}{1+e^{-t}}$ ;
  8. La *normalizzazione delle colonne di un database* serve per eguagliare gli intervalli di variabilità delle colonne;
  9. Per *ridurre il numero di falsi negativi* per un modello di classificazione dobbiamo massimizzare la sensibilità;
  10. Una partizione di un dataset si dice *stratificata* quando i campioni di ciascun sottoinsieme della partizione si trovano nello stesso ordine in cui compaiono nel dataset originale;
  11. È *possibile utilizzare KNN per la classificazione* se la classe in uscita ha più di due valori perché la funzione di decisione si basa sul valore di maggioranza, indipendente dal loro numero;
  12. La *K-fold cross validation* consiste nella separazione dei campioni in K gruppo distinti che si usano a rotazione per la validazione;
  13. Il *metodo di discesa lungo il gradiente* è un metodo per trovare un minimo locale di una funzione differenziale in più variabili;
- 
14. L'*Impurità di Gini* di una variabile casuale  $Y$  si *definisce come* la probabilità di errore nel prevedere un esito  $y \in Y$  se si sceglie un valore casuale  $\tilde{y}$  con la stessa distribuzione di probabilità;
  15. L'*entropia* di una variabile casuale  $Y$  dipende soltanto dai valori di probabilità;
  16. L'*impurità di Gini* di una variabile casuale  $Y$  dipende soltanto dal dominio di  $Y$ ;
  17. La *mediana della distribuzione* non risente molto della presenza di valori estremi (outliers) quindi è spesso opportuno utilizzarla per binarizzare una variabile continua, al posto della media;
  18. Se abbiamo una collezione di punti della forma  $(x, x_2)$ , con  $x \in [-1, 0]$  distribuito uniformemente, il *coefficiente di correlazione* fra le due coordinate vale  $\rho = 0$ , perché la relazione non è lineare;
  19. Il *calcolo dell'entropia* di una variabile casuale **non** è un algoritmo greedy;
  20. Ad ogni iterazione dell'algoritmo di clustering agglomerativo gerarchico su una matrice di distanze, *i due cluster da aggregare si scelgono sulla base del linkage criterion scelto*;
  21. Il *linkage criterion* **non** ha influenza sul bilanciamento del dendrogramma.