Algoritmi Avanzati

Emanuele Nardi

30 agosto 2018

Risposte alle domande a risposta multipla

Per il primo parziale

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. É 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].

   La funzione sigmoide non "decide" nulla, né determina valori di soglia, i quali devono semmai essere utilizzati a valle dell'applicazione della sigmoide;

7. La *funzione sigmoide* può essere definita come $\frac{1}{1+e^{-t}}$.

   É definita su tutti i reali;

8. La *normalizzazione delle colonne di un database* serve per eguagliare gli intervalli di variabilità delle colonne.

   Alcuni algoritmi sono sensibili a differenze eccessive fra gli ordini di grandezza dei diversi attributi; la normalizzazione serve a far variare tutti gli attributi all'interno di uno stesso intervallo. I valori negativi o nulli non sono, normalmente, un problema;

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. É 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. É possibile utilizzare l'algoritmo *KNN su problemi di regressione* calcolando la media delle y dei K elementi più vicini a quello incognito.

    Un esempio è fornito nelle dispense e considera la media pesata con pesi inversamente proporzionali alla distanza;

13. La *K-fold cross validation* consiste nella separazione dei campioni in K gruppo distinti che si usano a rotazione per la validazione;

14. Il *metodo di discesa lungo il gradiente* è un metodo per trovare un minimo locale di una funzione differenziale in più variabili reali.

    Si tratta di un algoritmo iterativo di approssimazione per trovare un minimo di una funzione. Non serve a trovare le derivate parziali (che si possono calcolare a tavolino o, in alternativa, stimare), e non ha alcuna connessione immediata con l'ottimizzazione di un classificatore KNN;

---

15. In una regressione polinomiale di terzo grado in una variabile vanno determinati *quattro coefficienti*.

    Un polinomio di terzo grado ha la forma $\beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0$, quindi servono quattro coefficienti;

16. Per valutare la precisione di un classificatore in una matrice di precisione vanno utilizzate le caselle dei veri positivi (TP) e dei falsi positivi (FP).

    La precisione si valuta considerando le sole risposte positive di un classificatore;

17. Per valutare la sensibilità (o recall) di un classificatore in una matrice di precisione vanno utilizzate le caselle dei veri positivi (TP) e i falsi negativi (FN).

    La sensibilità si valuta considerando soltanto i casi positivi del dataset;

18. Per valutare l'accuratezza di un classificatore vanno utilizzate *tutte* le caselle della matrice di confusione.

    Non è possibile calcolare l'accuratezza se non si conoscono tutte le caselle, visto che è necessario conoscere la numerosità del dataset;

19. É preferibile che gli insiemi di training e di validazione siano disgiunti perché siamo interessati a valutare le prestazioni del sistema su esempi non visti durante l'addestramento.

    La valutazione delle prestazioni richiede, semmai, che *le dimensioni dei due dataset siano massimizzate*, non minimizzate, ovviamente sotto il vincolo che i due insiemi siano disgiunti;

Per il secondo parziale

1. 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;

2. L'*impurità* di Gini di una variabile casuale Y dipende soltanto dal dominio di Y;

3. L'*entropia* di una variabile casuale Y dipende soltanto dai valori di probabilità;

4. 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;

5. 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;

6. Il *calcolo dell'entropia* di una variabile casuale **non** è un algoritmo greedy;

7. 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*;

8. Il *linkage criterion* **non** ha influenza sul bilanciamento del dendogramma.

---

9. In un *albero di decisione addestrato in base all'information gain*, la decisione attribuita a un nodo minimizza l'entropia attesa della variabile di output nei figli.

   Il fattore da valutare è sempre l'entropia della varibile di output, in quanto misura dell'incertezza del valore da prevedere;

10. In un *albero di decisione addestrato in base all'impurita di Gini*, la decisione attribuita a un nodo minimizza l'impurità attesa della variabile di output nei figli.

    L'obiettivo di un albero di decisione è di avere nodi puri, quindi di minimizzare l'impurità. Come nella domanda precedente, la variabile di cui ci interessa valutare l'incertezza è sempre l'output;

11. L'intervallo di variabilità dell'entropia di una distribuzione di probabilità discreta è $[0, +\infty)$.

    L'entropia di una variabile discreta non è mai negativa, e può assumere qualsiasi valore, a partire da 0 (esito certo). Per rendersi conto che il suo valore non è limitato, basta considerare la sua interpretazione come "numero di bit" necessari a rappresentare l'informazione;

12. L'intervallo di variabilità dell'impurità di Gini di una distribuzione di probabilità discreta è $[0, 1]$.

    L'impurita di Gini è una probabilità, quindi varia tra 0 e 1. In realtà, il valore 1 non è ottenibile;

13. L'intervallo di variabilità dell'indice di correlazione di Pearson fra due variabili casuali discrete è $[-1, 1]$.

    La correlazione è una covarianza normalizzata, e può assumere valori negativi;

14. Il parametro principale K dell'algoritmo K-means indica il numero di cluster in cui suddividere il dataset.

    K rappresenta il numero di centroidi o prototipi. Da non confondere, ovviamente, con l'omonimo parametro dell'algoritmo KNN. Il numero di iterazioni non è generalmente prefissato;

15. In un'iterazione dell'algoritmo di clustering agglomerativo gerarchico vengono uniti i cluster a distanza maggiore.

    I due cluster da unire sono sempre i più simili (o meno distanti), indipendentemente dal linkage criterion, che entra in gioco solo nella determinazione di queste distanze;

16. Sono necessarie $n - 1$ iterazioni per un'esecuzione completa dell'algoritmo di clustering gerarchico agglomerativo su un insieme di n elementi.

    Si parte da n cluster e ad ogni iterazione se ne uniscono due, riducendo di uno il numero complessivo. Si termina quando c'è un solo cluster;

17. Date due variabili casuali discrete X e Y, se la loro informazione mutua vale $I(X;Y) = 0$ possiamo dire che le due variabili sono indipendenti: la conoscenza dell'esito di X non ci dice nulla sull'esito di Y.

    Significa che l'entropia di X non varia se la si condiziona alla conoscenza di Y;

18. Date due variabili casuali discrete X e Y, se la loro informazione mutua vale $I(X;Y) = 1$ possiamo dire che le due variabili sono dipendenti: la conoscenza dell'esito di X riduce l'informazione necessaria a comunicare l'esito di Y.

    L'informazione mutua rappresenta la diminuzione dell'entropia di X quando si conosce Y. In questo caso la diminuzione c'è. L'entropia non misura dipendenze lineari. Si osservi che, dato che l'entropia può assumere qualunque valore positivo, una diminuzione pari a 1 non rappresenta necessariamente una dipendenza completa.