Progetto

martedì 10 novembre 2020 09:41

L'obiettivo che abbiamo per ora è quello di analizzare nell'insieme i dati che abbiamo, valutando tutte le relazioni che riusciamo ad identificare, giustificandone il significato in qualche modo.

- Che tipi di attrbuti abbiamo?
  Esaminando il dataset ci siamo resi conto di aver a che fare con attributi di tipo categorico, ordinale e numerico. Sono presenti solamente attributi discreti. Alcuni attributi categorici sono restituiti tramite numeri, ma essendo ognuno di essi collegato a categorie, tali attributi sono da considerarsi, appunto, categorici.

- *Com'è la qualità dei dati?*

- *Una visualizzazione dei dati aiuta nella comprensione?*
  Per quanto riguarda la visualizzazione dei dati, ci ritroviamo ad avere un numero di attributi molto alto. Certo, non tutti quanti hanno valore ai fini del progetto, ma bisogna adottare tecniche capaci di darci qualche insight sul dataframe plottano tutti gli attributi in 2 o 3 dimensioni. L'idea principale è quella di trovare una tecnica di visualizzazione che preservi nel miglior modo possibile la struttura dei dati.
  a. **Metodo PCA (Principal Component Reduction)** : Mappa lo spazio m-dimensionale in uno spazio q-dimensionale, con $q < m$. Lo scopo è variare il meno possibile la varianza nei dati degli attributi.
     Si basa sulla proiezione dei dati sulle componenti principali della matrice di covarianza (quindi su un sottospazio lineare), ovvero sui suoi vettori normalizzati aventi autovalore più grande. Bisogna però preparare i dati: traslarli in maniera tale che le medie siano nello zero ed eseguire la standardizzazione allo z-score
     $$x \rightarrow \frac{x - \mu_x}{\sigma_x}$$
     Quando proiettiamo i dati lungo le prime q componenti principali, la frazione (m è la dimensione del dataframe, ovvero il numero di attributi)
     $$\frac{\lambda_1 + .. + \lambda_q}{\lambda_1 + .. + \lambda_m}$$
     Ci dà la frazione della varianza che viene preservata durante la proiezione.
  b. **Metodo MDS (Multidimensional Scaling)**: E' un metodo che si basa sulla matrice delle distanze. Mappa le distanze dei dati in un sottospazio a dimensionalità minore in cui le distanze sono mantenute invariate il più possibile. Quello che praticamente si fa, quindi, è valutare le distanze di tutti i punti e costruire la matrice delle distanze. Tramite la minimizzazione della somma degli errori quadri si identificano le distanze nello spazio a più bassa dimensionalità.

- *Gli attributi sono correlati?*

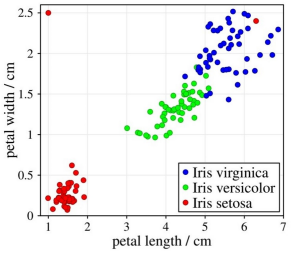
- *Ci sono outliers?*
  I **Box Plot** sono un primo strumento per identificare degli outliers per singoli attributi numerici. In un boxplot un outliers è tale se cade al di fuori della linea verticale che parte dal box e rappresentante 1.5 * Interquartile range (IQR = 3° quartile - 1° quartile).
  Un altro strumento utile per l'identificazione degli outliers per attributi categorici è **controllare la frequenza** con cui si presenta un dato: se estremamente bassa rispetto alle altre, è da considerarsi un outlier e quindi il dato va eliminato.
  Nel **caso multidimensionale** gli strumenti utilizzati sono gli **scatter plot, PCA o MDS plot e la Cluster Analysis**. In quest'ultimo caso gli outliers sono quei valori che non è possibile assegnare a nessun cluster.

- *Come vengono gestiti i missing values?*
  In generale si distinguono 3 tipi di missing values:
  a. **Missing values completamente random**: è il caso in cui la probabilità di avere un valore mancante è indipendente dal punto di vista condizionale dall'attributo di cui fa parte e gli altri attributi. In altre parole, si ha che i valori mancanti seguono la stessa distribuzione di probabilità dei valori presenti.
  b. **Missing values at random**: qui la situazione è leggermente diversa. La probabilità di avere un missing value dipende, sempre dal punto di vista condizionale, dai valori degli altri attributi. In questo caso, quindi, la distribuzione di probabilità di questi valori mancanti NON segue quella degli altri valori. ESEMPIO: Le batteria di una turbina posta come sensore alla velocità del vento si scaricano una volta ogni tanto, ma non vengono sostituite (con maggiore probabilità) durante i giorni di pioggia. Se gli altri attributi sono temperatura e pioggia, avrò che il valore mancante della velocità del vento è correlato e dipendente con temperatura bassa e presenza di pioggia
  c. **Non ignorable missing values**: in questi casi il fatto che il valore sia mancante è di per se significativo

  Come determinare la tipologia di missing value:
  - Sostiuire l'attributo che presenta missing values con uno nuovo X, binario, tale da contenere YES se il dato c'è e NO se non c'è
  - Costruire un classifier con l'attributo binario X come target e usare tutti gli altri attributi per la predizione della classe YES e NO
  - Determinare il misclassification rate, ovvero la porzione di data objects a cui non è stata assegnata la corretta classe dal classificatore.

  Nel caso di MAR, gli altri attributi non dovrebbero dare alcuna informazione aggiuntiva se X abbia un missing value o no. Quindi, il misclassification rate non dovrebbe differire molto dal puro indovinare, ovvero se ci sono 10% di missing value per l'attributo X, il misclassification rate del classificatore non dovrebbe essereminore del 10%. Se il misclassification rate del classificatore è significativamente migliore del puro indovinare, ciò è un indicatore che c'è correlazione fra missing values per l'attributo

Gli outlier possono essere relativi ad una singola classe o a più classi (all'intero dataframe in teoria). Nell'esempio sottostante c'è un outlier relativo alla classe Iris setosa ed uno relativo all'intero dataframe



A checklist for data understanding

- Determine the quality of the data. (e.g. syntactic accuracy)
- Find outliers. (e.g. using visualization techniques)
- Detect and examine missing values. Possible hidden by default values.
- Discover new or confirm expected dependencies or correlations between attributes.
- Check specific application dependent assumptions (e.g. the attribute follows a normal distribution)
- Compare statistics with the expected behavior.

A checklist for data understanding: Must Do

- Check the **distributions for each attribute**
  (unexpected properties like outliers, correct domains, correct medians)
- Check **correlations or dependencies** between pairs of attributes

X ed il valore per gli altri attributi.
Per quanto riguarda la gestione di tali missing values, si ha:

a. **Case deletion** è applicabile per i MCAR aventi un numero abbastanza grande di dati in maniera da non distorcerli troppo. Nel caso di MAR non è safe farlo.

b. **Imputation** I missing values possono essere sostituiti con altri valori, come ad esempio la media (per gli attributi numerici) o la moda (per gli attributi categorici). Sostituire con la media non affligge la media, ma la varianza si. E' possibile determinare un valore da sostituire che non vari la varianza, ma la media si. Il tutto dipende da quale indice statistico ci serve nello studio. Nel caso di MAR gli altri attributi possono dare degli hint per quelli mancanti. Modelli basati su quest'idea sono quelli di regressione e classification
Se l'attributo è numerico e continuo, è sensato sostituire il missing value con la media fra il precedente ed il successivo. Se l'attributo è categorico, si sostituisce con il valore con frequenza maggiore

a. **Explicit Value or Variable** Per gli attributi categorici è possibile adottare un approccio molto semplice introducendo una nuovo valore, MISSING, all'attributo. Tale approccio è sensato quando i dati mancanti sono di tipo non ignorabile, ovvero quando la loro stessa assenza può portare ad avere nuove informazioni sullo stesso dato. In questo caso, infatti, l'introduzione di una nuovo valore può esprimere qualcosa che non è possibile recuperare dagli altri attributi. Nel caso di MAR e MCAR non ha senso aggiungere una nuova classe: introducerebbe semplicemente complessità.