

Lezione 12 15/05/2024

Accuratezza

Funziona bene per alcuni tipi di dato e male per altri.

	Accuracy	Completeness	Currency	Consistency
Alphanumeric value	X			
Numerical value				
Tuple	X			
Relation	X			

è la vicinanza tra il valore nel db e il valore reale.

- if the name of a person is “John”
- the value $v = \text{“John”}$ is accurate,
- while the value $v1 = \text{“Jhn”}$ is inaccurate.

L'accuratezza sintattica è definita quando il vocabolario di riferimento dei valori è conosciuto, ed è il grado a cui i valori correttamente rappresentano i valori di dominio del vocabolario.

L'accuratezza semantica è definita come il grado di accuratezza della rappresentazione a livello di significato nel mondo reale.

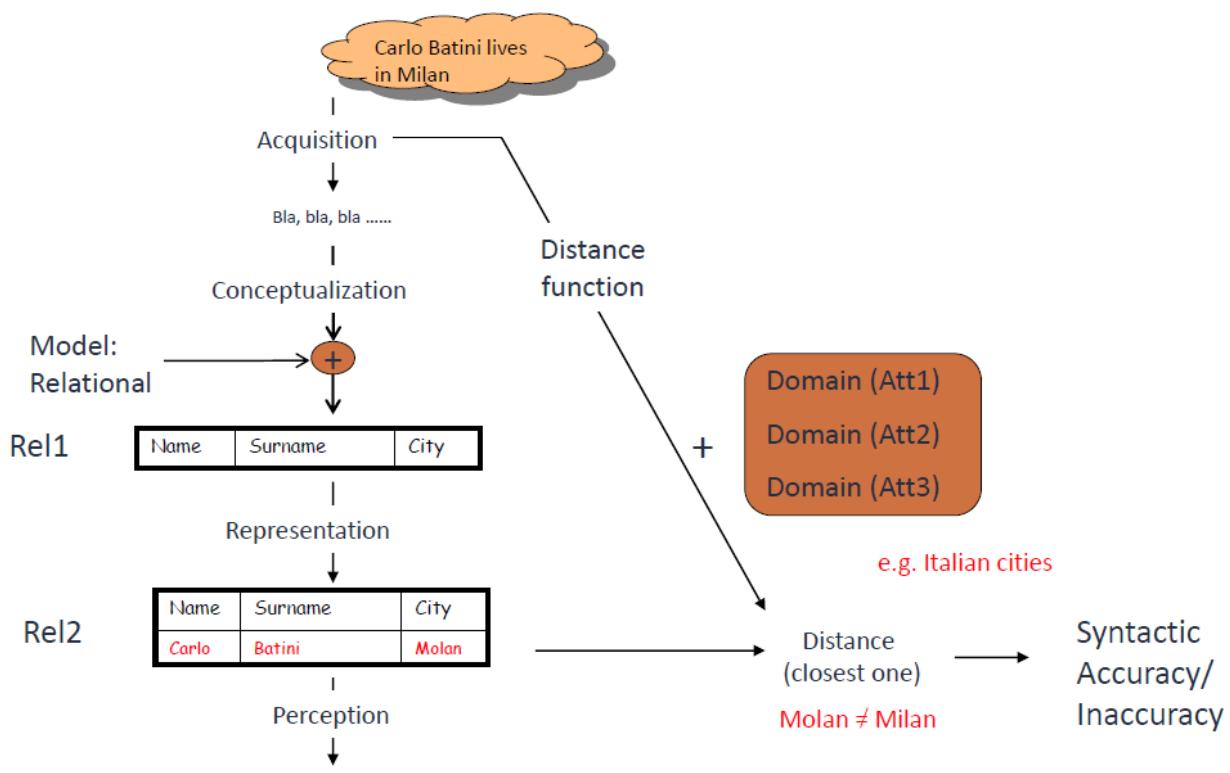
Esempio

1	Miroslav	Konecny	1978	7	10	Street	St. John	49	Prage	412776	null
2	Mrtin	Necasky	1982	7	8	Sq.	Wien	null	Bratislava	101278	Slovakia
3	Miroslav	Konecny	null	null	null	Str.	SaintJon	49	Prague	412776	null
4	Calo	Btini	1949	June	7	Street	Dessie	15	Rome	0019h	Italy
5	Miroslav	Knecy	1978	7	null	Sq.	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	September	7	Street	Sesto	null	Mian	20...	Ital
7	Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy

Per esempio Mrtin non è un dato sintatticamente corretto, mentre Anisa non sappiamo se è un nome reale o un errore, e quindi si parla di accuratezza semantica.

Metriche per l'accuratezza sintattica

è possibile fare qualcosa di automatico.



Calcoliamo una funzione di distanza per capire quale potrebbe essere l'errore, e quale potrebbe essere la città che nel mio elenco di città reali è la più vicina.

Quindi se il dato non è sintatticamente troppo lontano si può sistemare in modo automatico con questo metodo.

Se un dato non è sintatticamente corretto non lo sarà anche semanticamente, mentre un dato semanticamente corretto può essere non corretto sintatticamente.

Con i nomi è un po' più complicato.

Bisognerà come prima avere un elenco di tutti i possibili nomi per calcolare la vicinanza.

From which Italian name does it come → «Carl»?

Differ both in one letter....

	MARIO	% sul totale dei maschi
1999	1.384	0,52
2000	1.374	0,50
2001	1.354	0,50
2002	1.244	0,40
2003	1.273	0,47
2004	1.290	0,45
2005	1.199	0,43
2006	1.248	0,44
2007	1.104	0,39
2008	1.150	0,39
2009	1.158	0,40
2010	1.128	0,40
2011	1.001	0,37
2012	1.095	0,47
2013	1.021	0,40
2014	983	0,37
2015	1.005	0,41

	MARIA	
	valore assoluto	% sul totale delle femmine
1999	2.315	0,92
2000	2.391	0,93
2001	2.146	0,84
2002	1.904	0,74
2003	1.921	0,74
2004	1.902	0,71
2005	1.962	0,74
2006	1.933	0,72
2007	1.828	0,68
2008	1.824	0,66
2009	1.751	0,64
2010	1.754	0,66
2011	1.648	0,64
2012	1.637	0,65
2013	1.702	0,70
2014	1.687	0,70
2015	1.597	0,68

From which Italian name comes → «Mari» ?

In absence of any other information,
Mario → 42%, Maria → 58%

Si può usare un metodo probabilistico.

Nel caso di dati strutturati, come stringhe di città lunghe, al posto di confrontare i singoli caratteri, si lavora a token.

- Based on the alphanumeric strings
 - «Santa Margherita Ligure»
- Based on their structure in terms of items (tokens)
 - «Santa» «Margherita» «Ligure»

Per le stringhe si utilizza la distanza di edit oppure altre funzioni simili.

$$\begin{aligned} \text{UED(Maro, Mario)} &= 1 \\ \text{UED(Maro, Maria)} &= 2 \\ \text{UED(Maro, Margherita)} &= 7 \end{aligned}$$

...

Esiste anche la versione normalizzata

$$ED_{\text{norm}}(v1, v2) = 1 - ED(v1, v2)/n$$

Completezza

La completezza si applica a tutti i valori, può essere definito come la copertura con la quale il fenomeno osservato è rappresentato nell'insieme di dati. Quindi per esempio dove c'è un valore null o un valore fuori dominio (pericoloso, tipo età -1)

Quindi basta andare a contare il numero di valori nulli per riga (completezza di tupla), colonna (completezza di attributo o colonna) o tabella (completezza di tabella).

Nelle precedenti definizioni, abbiamo assunto una ipotesi di mondo chiuso: tutto cio' che e' rappresentato nella BD e' vero, tutto il resto e' falso. Questa e' la tipica ipotesi che si fa nei DBMS.

Una ipotesi alternativa e' quella di mondo aperto: di tutto cio' che non e' rappresentato non si sa nulla. In questo caso introduciamo la **object completeness**, che tiene tenere conto del fatto che gli oggetti rappresentabili sono piu' delle tuple della tabella, e di tale cardinalita' serve una stima indiretta

Esempi

Impiegato

Cognome	Eta'	Citta'Nascita
Rossi	null	null
Verdi	35	Roma
Neri	null	Milano

Completezza degli attributi:

Cognome	100%
Eta'	33%
Citta'Nascita	66%

Impiegato

Cognome	Eta'	Citta'Nascita
Rossi	null	null
Verdi	35	Roma
Neri	null	Milano

Nella ipotesi che gli impiegati siano 4

Completezza degli oggetti: 75 %

Currency

La currency misura con quale rapidità i dati sono aggiornati (rispetto al corrispondente fenomeno del mondo reale). È una metrica difficile da misurare in alcuni casi.

La currency può essere definita come differenza tra tempo di arrivo alla organizzazione e tempo in cui è effettuato l'aggiornamento. È misurabile se c'è un log degli arrivi e degli update.

Tempestività

La tempestività misura quanto i dati sono aggiornati rispetto a un particolare processo (o ai processi) che li utilizza.

Esempio: l'orario delle lezioni e delle aule deve essere disponibile prima dell'inizio dei corsi.

La tempestività, al contrario della currency, è dipendente dal processo, ed è associata al momento temporale in cui deve essere disponibile per il processo che utilizza il dato.

Possono esistere dati con elevata currency, ma ormai obsoleti per il processo che li usa. Se l'orario è stato prodotto fuori tempo massimo ed è stato caricato subito dopo essere stato prodotto è current ma non è tempestivo.

Consistenza

Ha due significati:

- Consistenza dei dati con i vincoli di integrità definiti sullo schema. Es: CAP deve essere consistente con Città.
- Consistenza delle diverse rappresentazioni di uno stesso oggetto della realtà presenti nella base di dati. Es: L'indirizzo deve essere rappresentato con lo stesso formato in tutte le basi di dati in cui è definito.

I vincoli di consistenza (o business rules) possono includere:

- un singolo attributo (i valori dell'attributo devono rientrare nel dominio {1..10});
- più di un attributo (CAP e città non devono essere in conflitto);

- Attributi in più relazioni (es integrità referenziale)
- Essere espressi in termini di probabilità: Esempio: età > 70 anni e colore capelli = nero → probabilità = 0.001

Accessibilità

Esprime la capacità di un utente di accedere ai dati a partire dalla propria cultura, stato fisico e psichico e dalla tecnologie disponibili. Quindi per esempio la lingua in cui è scritto, oppure renderli accessibili per persone con disabilità.

Tradeoff tra qualità

Consistenza e completezza nel modello relazionale possono essere non conciliabili quando si voglia rispettare la integrità referenziale.

Nel web, per "arrivare primi" con una notizia si privilegia in molti casi la tempestività rispetto a accuratezza e completezza.

Quality assessment and improvement

Per fare l'improvement, ci sono due strategie: una guidata dai dati e una guidata da un processo.

La strategia data driven va a pulire il dataset in base a delle metriche scelte, facendo comparazione tra i dati.

La strategia process driven è fatta da quegli approcci che cercano di migliorare il dato agendo sul processo che crea o modifica i dati, inserendo dei controlli di qualità nelle varie fasi.

Si può anche fare redesign del processo per andare a rimuovere le cause della qualità bassa dei dati, introducendo delle nuove attività che producono dati migliori. Se il processo di redesign è radicale, allora è detto business process reengineering.

Le tecniche process driven sono quelle più efficaci

Data quality for machine learning

Si vogliono evitare noisy labels ovvero accuratezza semantica e consistenza, gli outliers ovvero dati che non sono nel dominio di riferimento, overlapping, problemi

di inconsistenza e missing values.

Noisy label

Possibili motivazioni legate ad errate etichettature: informazioni insufficienti date all'esperto, errori nella etichettatura, soggettività nell'etichettatura, problemi di comunicazione, codifica. Gli effetti sono: riduzione delle performance, possibili errori nella fase di feature selection, e in fase di produzione i dati reali potrebbero contraddirre quelli di training.

Ci sono due approcci per risolvere:

- Approcci algoritmici, ovvero progettare algoritmi che siano robusti rispetto agli errori, però sono difficili da progettare e non sono facilmente portabili in altri contesti.
- Approcci ai dati, andando ad eliminare le etichette sbagliate prima di eseguire i task di machine learning. Questo è quindi indipendente dal metodo di ML utilizzato.

Gli approcci filter based sono quelli che usano un filtro per identificare eventuali errori nelle labels

Outlier

Un outlier è una osservazione che devia molto dalle altre e che solleva il sospetto che sia stata generata da un differente meccanismo.

Può essere causato da data entry errate, da errori di misura dello strumento, errori di campionamento, errori di procedure, oppure da origini naturali (esempio c'è un incendio e quindi il sensore di calore registra valori alti).

Gli algoritmi di ML sono sensibili alle distribuzioni dei valori e al loro range.

Gli outlier possono portare a errori nella definizione del modello e ridurre l'accuratezza del modello.

Missing

I missing value rappresentano i casi in cui il valore di una variabile non è disponibile. Possono arrivare da errori di misura, del sensore, di comunicazione del dato, oppure da un questionario incompleto.

Però gli algoritmi di ML di solito non accettano valori nulli. Si può quindi sostituirli cancellandoli, oppure inserendo altri valori già presenti nel dataset, si può assegnare la media, moda etc, oppure si possono usare dei modelli di regressione o classificazione per predire il valore mancante.

Anche in questo caso esistono dei tradeoff fra le varie dimensioni di qualità da considerare e le relative metriche. Da diversi studi sembra che l'ordine con cui si realizzano le attività di improvement impatti la qualità dei risultati dei task di ML.