

# Ricerca di outlier

Prof. Matteo Golfarelli

Alma Mater Studiorum - Università di Bologna

## Ricerca di Anomalie/Outlier

- Cosa sono gli outlier?
  - ✓ L'insieme di dati che sono considerevolmente differenti dalla maggior parte dei dati
- Formulazioni alternative del problema di ricerca di outlier:
  - ✓ Dato un database  $D$ , trovare tutti i punti  $x \in D$  con un *livello di anomalia* superiore a una soglia  $t$
  - ✓ Dato un database  $D$ , trovare tutti i punti  $x \in D$  che presentano i top- $n$  livelli di anomalia
  - ✓ Dato un database  $D$ , contenente principalmente punti "normali" (ma non classificati come tali), e un punto di test  $x$ , calcolare il livello di anomalia per  $x$  rispetto a  $D$
- Applicazioni:
  - ✓ Identificazioni di frodi nell'uso di carte di credito, intrusioni in una rete, identificazione di errori, pulizia di dati



## Cause delle anomalie

- **Dati da classi differenti:** un oggetto può risultare differente poiché appartenente a una diversa classe
  - ✓ Il truffatore che ha rubato una carta di credito segue un pattern di acquisto diverso da quello del legittimo proprietario
  - ✓ Anomalie di questo tipo sono spesso l'oggetto della ricerca
- **Variazioni naturali:** molti fenomeni possono essere modellati con distribuzioni probabilistiche in cui esiste, anche se molto ridotta, la probabilità che si verifichi un fenomeno con caratteristiche molto differenti dagli altri
  - ✓ Una persona alta 210cm non è anomala perché appartiene a una classe diversa, ma perché la sua caratteristica altezza assume un valore "estremo" rispetto alla popolazione
  - ✓ Anomalie di questo tipo sono spesso di interesse e oggetto di studio
- **Errori di misurazione:** dovuti a errori umani o dei dispositivi
  - ✓ La ricerca di questo tipo di anomalia è finalizzata all'esclusione dal data set dato che rappresenta un rumore che può pregiudicare la qualità dei risultati di analisi

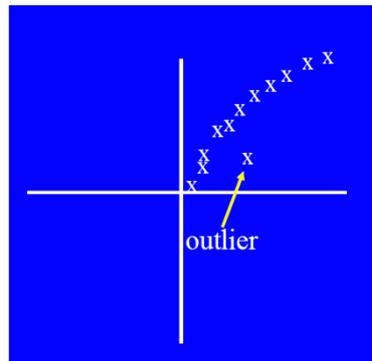
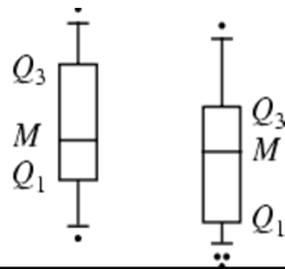


## Ricerca di Anomalie/Outlier

- Elementi di complessità
  - ✓ Quanti outlier ci sono nei dati?
  - ✓ Il problema richiede normalmente tecniche non supervisionate
    - La validazione è spesso molto complessa come nel caso del clustering
- Assunzioni:
  - ✓ Il numero delle osservazioni "normali" è largamente superiore a quelle "anormali"
- Approccio generale
  - ✓ Costruisci un profilo del comportamento "normale"
    - Un profilo può essere definito tramite pattern o statistiche riassuntive della popolazione
  - ✓ Utilizza il profilo "normale" per individuare le anomalie

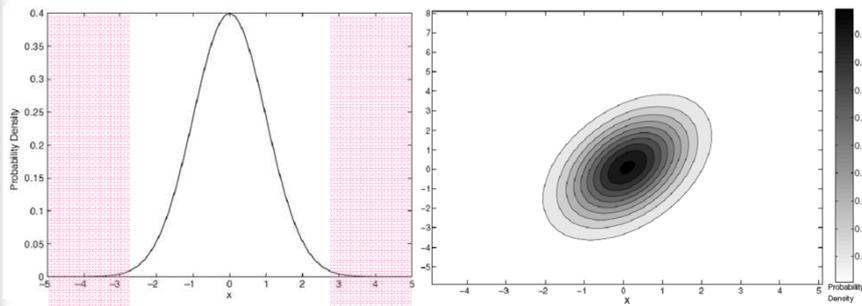
## Approcci grafici

- Analisi manuale dei dati svolta con il supporto di opportune tecniche di visualizzazione
  - ✓ Boxplot (1-D), grafici a dispersione (2-D), ecc.
- Limiti
  - ✓ Può richiedere molto tempo
  - ✓ Soggettivo
  - ✓ Di difficile applicazione a dati multidimensionali



## Approcci statistici

- Assumendo l'esistenza di un modello parametrico che descrive la distribuzione dei dati (es. Distribuzione gaussiana)
- Si esegue un test statistico in cui si fissano
  - ✓ I parametri della distribuzione (es. media e StdDev)
  - ✓ Il numero atteso di outlier o equivalentemente un valore di soglia di probabilità
  - ✓ I punti che presentano una probabilità sufficientemente ridotta sono considerati outlier





## Limiti degli approcci statistici

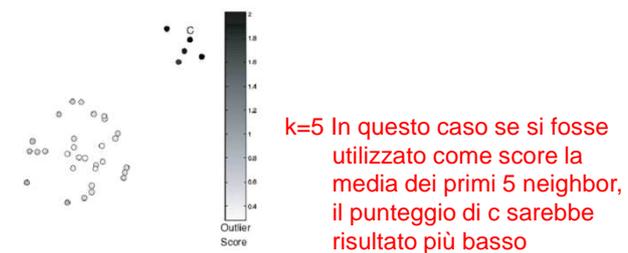
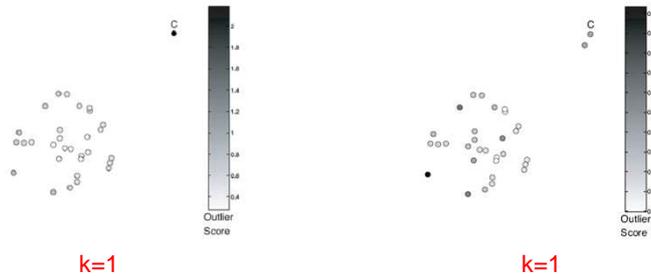
- La maggioranza dei test è per distribuzioni mono dimensionali
- In molti casi le distribuzioni non sono note
- Per dati multidimensionali con elevata dimensionalità può essere molto difficile stimare la distribuzione dei dati con accuratezza



## Approcci basati sulla distanza

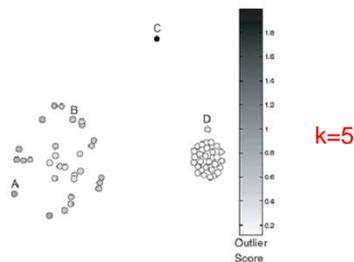
- I dati sono rappresentati da vettori di caratteristiche
- Un dato rappresenta un outlier se è distante dalla maggioranza degli altri punti
  - ✓ Questa definizione è più facilmente applicabile a dataset reali dato che è più semplice identificare un'appropriata misura di prossimità/distanza piuttosto che una precisa distribuzione dei dati
- **Def. di outlier basato sulla distanza:** il punteggio di outlier di un punto  $x$  è calcolato come distanza del suo  $k$ -nearest neighbor
  - ✓ Il punteggio dipende dal valore di  $k$ 
    - Se  $k$  è troppo piccolo un insieme di outlier vicini possono determinare un punteggio di outlier basso ed essere considerati un cluster normale
    - Se  $k$  è troppo grande al contrario tutti i punti di un cluster normale possono diventare outlier
- Per rendere l'approccio più robusto rispetto al valore di  $k$  conviene utilizzare come score la media della distanza del punto dai primi  $k$ -nearest neighbor

## Approcci basati sulla distanza



## Approcci basati sulla distanza

- Hanno complessità  $O(N^2)$  con  $N$  numero dei punti nel dataset
- Sono sensibile alla scelta di  $k$
- Non possono gestire dataset con densità variabile dei dati



- Il punto C viene correttamente individuato, ma come deve essere considerato il punto D?

## Approcci basati sulla densità

- **Def. di outlier basata sulla densità:** un outlier è un elemento del data set posizionato in una zona a bassissima densità
  - ✓ Gli approcci basati sulla densità sono simili a quelli basati sulla distanza dato che la densità è definita in termini di distanza di un punto dai suoi vicini
- La **densità** di un elemento  $\mathbf{x}$  di un dataset  $D$  è calcolata come l'inverso della media delle distanze dai suoi  $k$ -nearest neighbor

$$\text{density}(\mathbf{x}, k) = \left( \frac{\sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{distance}(\mathbf{x}, \mathbf{y})}{|N(\mathbf{x}, k)|} \right)^{-1}$$

dove  $N(\mathbf{x}, k)$  è l'insieme dei  $k$ -nearest neighbor di  $\mathbf{x}$  e  $|\dots|$  denota la cardinalità dell'insieme

- ✓ In alternativa potrebbe essere utilizzato il concetto di densità utilizzato in DBSCAN

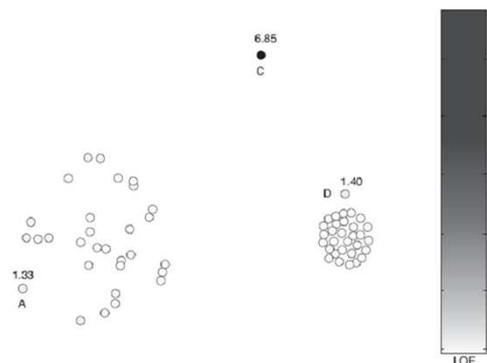
## Approcci basati sulla densità

- La precedente definizione di densità presenta gli stessi limiti delle tecniche basate sul concetto di distanza/prossimità: non permette di gestire dataset con densità eterogenee
- Per superare questo limite è necessario definire il concetto di **densità relativa** definita in funzione della densità dei suoi vicini

$$\text{AVGRelDensity}(\mathbf{x}, k) = \frac{\text{density}(\mathbf{x}, k)}{\frac{1}{|N(\mathbf{x}, k)|} \sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{density}(\mathbf{y}, k)}$$

## Approcci basati sulla densità

- Utilizzando il concetto di densità relativa è possibile identificare efficacemente gli outlier anche per data set con densità variabile
- La densità relativa è utilizzata per misurare localmente la tendenza di un punto a essere un outlier **Local Outlier Factor** (LOF)



## Outlier in spazi con elevata dimensionalità

- Per dataset D con elevata dimensionalità  $d$ , i dati sono sparsi e le definizioni di prossimità, distanza e densità diventano poco significativi
  - ✓ Le distanze tra i punti diventano molto simili e distanti
- Conviene quindi utilizzare tecniche diverse per la ricerca di outlier basate sulla proiezione dei punti su spazi a dimensionalità minori
  - ✓ Un punto con dimensionalità  $d$  è un outlier se esiste una sua proiezione in uno spazio con dimensionalità  $k < d$ , in cui il punto appartiene a una regione con densità fortemente bassa
- Possiamo quindi tentare di determinare quali regioni dei  $k$ -spazi presentino tali ridotte densità e quindi identificare come outlier quei record che vengono proiettati su quelle regioni

## Outlier in spazi con elevata dimensionalità

- L'individuazione di regioni a bassa densità avviene comparando la densità effettiva della regione con la densità che dovrebbe avere la regione assumendo distribuzione uniforme dei punti nello spazio
- Dividi i valori di ognuno dei  $d$  attributi in intervalli equi-eight  $\phi$ 
  - ✓ A ogni intervallo sarà associata una frazione  $f = 1/\phi$  del numero totale dei record
- Consideriamo ora una qualsiasi proiezione  $k$ -dimensionale ottenuta scegliendo  $k$  dei  $d$  attributi
  - ✓ La suddivisione dei valori degli attributi in intervalli determina delle celle  $k$ -dimensionali
  - ✓ Ipotizzando indipendenza tra i diversi attributi e distribuzione uniforme dei dati a ogni cella sarà associata una frazione  $f^k$  dei record
  - ✓ Se il dataset contiene  $N$  record, ogni cella sarà associata in media a  $N \cdot f^k$  record
  - ✓ Il numero medio di record nella cella ( $N \cdot f^k$ ), per il teorema del limite centrale, segue una distribuzione gaussiana con DevStd

$$\sqrt{N \cdot f^k \cdot (1 - f^k)}$$

## Outlier in spazi con elevata dimensionalità

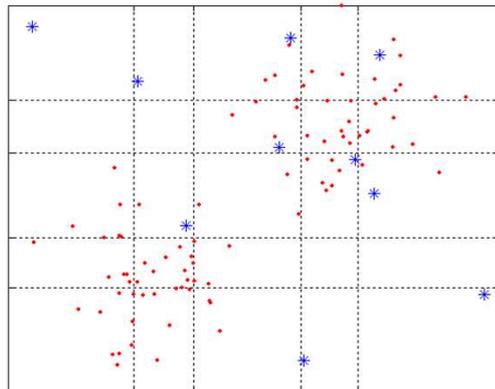
- Consideriamo ora una specifica proiezione  $k$ -dimensionale che utilizza un insieme  $\mathbf{K}$  di attributi ( $|\mathbf{K}|=k$ ).
  - ✓ Dato che nella pratica i dati non sono uniformemente distribuiti a ogni cella  $\mathbf{C}$  della proiezione sarà associata un certo numero di record  $n(\mathbf{C})$ .
  - ✓ Ogni cella  $\mathbf{C}$  è definita da una coppia di coordinate per ognuna delle  $k$  dimensioni coinvolte nella proiezione
- La sparsità della cella  $\mathbf{C}$  può essere definita come:

$$S(\mathbf{C}) = \frac{n(\mathbf{C}) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$

- ✓ Valori negativi di sparsità indicano celle con un numero di punti minore di quello atteso
- ✓ Dividendo per la StdDev si pesano diversamente gli scostamenti dai valori attesi in base all'ampiezza della distribuzione normale

## Outlier in spazi con elevata dimensionalità

- Proiettando 100 record multidimensionali su uno spazio 2-dimensionale si ottiene il seguente risultato
  - ✓ I punti in azzurro sono i punti etichettati come outlier da un esperto del dominio



$N=100, \phi = 5$   
 $f = 1/5 = 0.2$   
 $N \times f^2 = 4$