DATA STREAM

I data stream sono una tipologia di dati illimitati, ordinati secondo una sequenza di istanza. Può essere visto come una sequenza $S = \{x_1 \dots x_N\}$ dove x_i è l'iesima istanza di un array di feature di dimensione d e con $N \to \infty$.

Le tecniche tradizionali di Data Mining hanno bisogno:

1) Dell'intero dataset finito.
2) Spesso scansioni multiple sono necessarie
3) Accesso casuale alla istanza
4) Fase di addestramento dispendiosa.

Tutte cose che per i data stream non vanno bene, perché siamo impossibilitati ad assecondare tali richieste. Però nel mondo moderno, sempre più realtà hanno bisogno di avere a che fare con data stream.

Necessità dei data stream

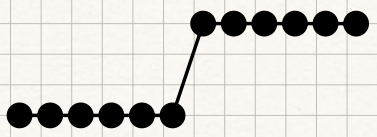
1) Single Pass
2) Memoria limitata
3) Real-Time processing

CONCEPT DRIFT

È quando si manifesta un cambiamento improvviso delle proprietà statistiche dei dati osservati nel tempo.

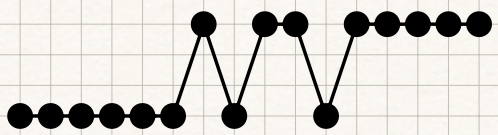
1) Sudden concept drift

Cambiamento unico a un certo momento fra due istanze consecutive. Dopo il cambiamento le istanze successive seguiranno la nuova tendenza



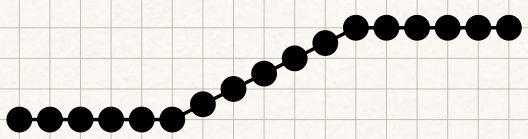
2) Gradual Concept Drift

Questa volta il cambiamento non è repentino ma graduale, iniziando con qualche istanza che comincia a cambiare trend, per poi trovare una stabilizzazione



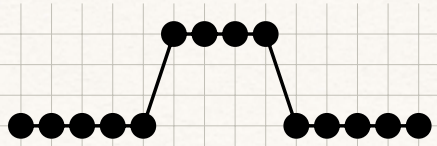
3) Incremental Concept Drift

Le istanze gradualmente cambiano trend, step by step.



4) Recurring Concept Drift

Le istanze balzano fra due classi molte volte nel tempo senza che nessuna classe sparisca



STRUTTURARE I DATI

Dato che non è possibile salvare e gestire una tale mole di dati, si cerca di usare strutture dati che riepilogano i dati da analizzare

1) FEATURE VECTORS: riassunto dei dati
2) PROTOTYPE ARRAY: tenere campioni significativi
3) CORESET TREES: riassunto ma in un albero
4) GRIDS: Usare la densità dei dati nel feature-space

THE WINDOW MODEL

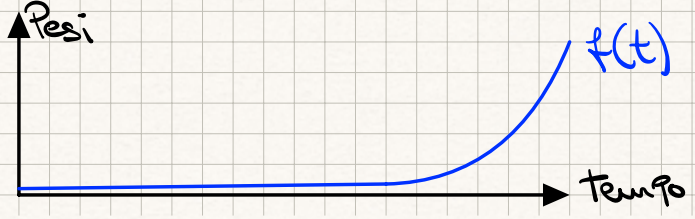
È molto meglio analizzare i dati recenti rispetto a tutti i dati.

1) DAMPED WINDOW MODEL

Pesa i dati in modo da preferire i più recenti.
Spesso usa un modello con funzione di decadimento che fanno calare i pesi dei dati più vecchi:

$$f(t) = 2^{-\lambda t}$$

Dove λ è il parametro di decadimento, più è alto più il decadimento dei dati passati è importante.



2) LANDMARK WINDOW MODEL

Tutti i dati fra due punti di Landmark hanno uguale peso. Window consecutive non si sovrappongono.
Ogni finestra ha peso dei dati diverso.

3) SLIDING WINDOW MODEL

Le finestre si sovrappongono, e scambiano un'istanza a ogni step, in particolare l'istanza più vecchia viene abbandonata per fare spazio a una più nuova.