Suggerimenti per l'approccio all'analisi dei dati multivariati

Struttura della presentazione della tesina

- **Titolo**
- **Sommario**
- **Obiettivi**
- **Contestualizzazione dei dati**
  - performance della letteratura e/o problemi riscontrati e non ancora risolti in letteratura
- **Analisi dati**
  - Eventuale descrizione (sintetica) delle tecniche di analisi dati se nuove per l'audience..
  - Rappresentazione dei risultati dati in maniera semplice
    - Se c'e' analisi esplorativa
    - Modello di regressione/ pattern recognition
  - Discussione dei risultati (questa puo' essere eventualmente inserita nel punto precedente) e loro confronto con la letteratura
- **Conlusioni**
  - Evidenziare gli obiettivi
  - Prevedere gli sviluppi futuri

Definizione degli obbiettivi

- Il primo passo è la definizione degli obbiettivi.
  - Qual è l'obbiettivo della sperimentazione i cui dati dovete analizzare.
    - Analisi esplorativa
    - Necessità di un modello di regressione
    - Pattern recognition

Prima di partire (I)

- Raccolta di tutte le informazioni disponibili sui dati
  - Come è stata organizzata la sperimentazione
    - Come sono state fatte le misure (protocollo)
    - Che tipo di variabili esterne hanno preso in considerazione
  - Rendere tutte le informazioni qualitative in formato numerico
  - Formato elettronico dei dati (file testo o Excel).

Prima di partire…(II)

- Necessità di contestualizzare i dati
  - Che tipo di dati sono
    - dati sensoriali, economici, etc.
    - Provengono da un unico strumento o da più strumenti.
      - Sono tutti dello stesso tipo o dimensionalmente differenti (kg, m, Pa,…)
      - Conoscenza degli strumenti o dei sistemi utilizzati nella sperimentazione
  - Ricerca bibliografica!
    - Verificare se un problema simile è stato trattato in letteratura.
      - Che tipo di analisi è stata fatta
      - Quali erano gli obbiettivi dello studio
      - Dove il problema è eventualmente simile e dove differente.

Definizione degli obbiettivi

- Analisi Esplorativa
  - Correlazioni tra le variabili
  - Capacità dei dati di raggrupparsi in classi
  - Variabili identificative di classi specifiche
  - Set minimo di variabili per la rappresentazione e variabili non fondamentali alla rappresentazione.
  - Identificazioni e giustificazione di possibili Outliers.
  - Considerazioni sulla bontà del dataset analizzato:
    - Il numero di dati è sufficiente per ottenere un modello statisticamente significativo?
    - Capire se ed eventualmente come deve essere impostata una nuova campagna di dati.

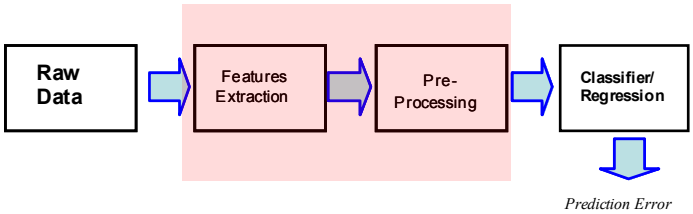
Definizione degli obbiettivi

- Modelli di regressione
  - Identificazioni degli obbiettivi
  - Scelta della tecnica di regressione (LR, MLR,PCR,PLS)
  - Scelta della tecnica di validazione del modello (LOOCV, K-fold, ….)
  - Riduzione delle variabili in ingresso
  - Identificazioni e giustificazione di possibili Outliers.
  - Validità del modello nel tempo e in diverse condizioni di misura
    - Il numero di dati è sufficiente per ottenere un modello statisticamente significativo?
    - Capire se ed eventualmente come deve essere impostata una nuova campagna di dati.

Definizioni degli obbiettivi

- Pattern Recognition:
  - Definizioni delle classi e delle percentuali minime di corretta classificazione.
    - Definizioni delle condizioni al contorno : entro quanto tempo
    - Qual è la classificazione critica (approccio dicotomico : possibile suddivisione in più macro classi).
  - Scelta del tipo del tipo di cluster analysis (lineare o non lineare , rete neurale , fuzzy …)
  - Metodo di validazione.
  - Identificazione degli outliers.
  - Confronto con altre tecniche
  - Validità del modello nel tempo e in diverse condizioni di misura
    - Il numero di dati è sufficiente per ottenere un modello statisticamente significativo?
    - Capire se ed eventualmente come deve essere impostata una nuova campagna di dati.

Quando i “raw data” non bastano…



Molto spesso i dati grezzi, cioè così come vengono dati non sono sufficienti ad ottenere l’obbiettivo fissato. Le possibili cause :

- Rumore troppo grande;
- Dai dati grezzi la tecnica di regressione non riesce ad estrarre l’informazione necessaria per ottenere un modello che dia delle buone prestazioni;

In questo caso diventa necessario almeno uno dei due step (Features Extraction o PreProcessing)