Università degli Studi di Padova
Corso di Laurea Magistrale in *Informatica*
a.a. 2018/2019

Data Mining

Docenti: A. Canale, M. Cattelan, D. Risso

Alcune indicazioni per l'analisi di un dataset

Di seguito si riporta una serie di indicazioni per l'analisi di un dataset, che riassumono quanto visto durante le settimane di lezione. Come ricordato più volte, **il modello giusto non esiste**: siamo alla ricerca di un buon modello o di modelli buoni basati su approcci anche molto diversi tra loro (un approccio semiparametrico ed un approccio parametrico, ad esempio). Puntiamo ad un modello semplice, facilmente interpretabile, con buone capacità previsive. Persone diverse che analizzano gli stessi dati possono giungere a modelli diversi, che avranno molti aspetti di similarità, ma anche differenze sostanziali.

I **punti fermi** su cui è bene non commettere errori che pregiudicano l'analisi sono, invece, i seguenti: applicare un modello di regressione quando si ricerca una classificazione e viceversa, confrontare modelli non annidati con tecniche che richiedono annidamento, non trattare le variabili qualitative come fattori, non rispettare il principio di gerarchia, mal interpretare l'analisi dei residui, mal interpretare le tecniche di selezione e confronto tra modelli, ...

Le indicazioni seguenti sono molto generiche: **l'analisi di ogni dataset è un processo a sè** che può seguire strade diverse.

1. Individuare se si tratta di una regressione (Y continua) o di una classificazione (Y 0/1).
2. Ci sono dati mancanti? Se si, e se non sono troppi, eliminarli dal dataset.
3. Iniziare l'analisi con alcune **ispezioni grafiche** dei dati.

    - Distribuzione di Y. Nel caso di Y continua, la valutazione della sua distribuzione procede tipicamente tramite boxplot e istogramma. Se l'indicazione grafica è quella di una distribuzione non normale si può valutare se una trasformata (logaritmica?) aiuta a recuperare la normalità. Il boxplot della variabile `box` nel dataset `film` mostra che la variabile logaritmica è adatta: veniva infatti suggerita nella consegna dell'esempio di prova pratica. Nel caso di Y categoriale si può valutare l'omogeneità o meno dei gruppi in termini di numerosità delle osservazioni tramite la costruzione di una tabella.
    - Si passa poi alle relazioni tra Y e le potenziali esplicative.
        - Se Y è continua si utilizzano diagrammi di dispersione rispetto alle esplicative continue (`plot(cmngsoon, box)` nel dataset `film` o `pairs(dati)`) oppure boxplot rispetto alle esplicative categoriali (`boxplot(box~animated)` nel dataset `film`). Un diagramma di dispersione tra Y ed una esplicativa continua può essere costruito evidenziando i punti associati ad ogni livello di una variabile categoriale per valutare se esistano interazioni tra le esplicative. Ad esempio, il diagramma di dispersione tra `box` e `cmngsoon` evidenziando i punti `animated=FALSE` e `ANIMATED=TRUE` può suggerire una interazione tra le variabili esplicative nel dataset `film`.
        - Se Y è categoriale si utilizzano grafici boxplot per le eventuali relazioni con esplicative continue (`boxplot(cmngsoon~box.categorica)` nel dataset `film`).
        - Le relazioni tra Y categoriale e altre esplicative categoriali si valutano tramite costruzione di tabelle o mosaicplot (`mosaicplot(table(box.categorica,action))` nel dataset `film`).

- Si tenga sempre presente che i grafici forniscono delle informazioni iniziali per suggerire come le esplicative possono entrare nel modello (polinomio? interazione? modo non lineare?). Sarà poi l'analisi del modello (p-value, anova, residui, ...) ad aiutarci a stabilire se le variabili saranno significative o meno, se il comportamento sarà diverso dalle attese grafiche e così via. Il grafico è una visione in due dimensioni di una relazione che non può tradurre tutte le relazioni esistenti tra le variabili. Quella che può sembrare una relazione parabolica tra Y (ad esempio `box`) e X (ad esempio `cmngsoon`) potrebbe rivelarsi solo una relazione lineare se altre esplicative catturano parte della variabilità espressa da X.

4. Si inizia la **costruzione del modello**, un processo (a volte lungo) che va avanti a passi, che richiede valutazioni grafiche ed analitiche continue, semplificazioni, aggiustamenti, ...

   - Si può partire da un modello lineare o logistico che preveda l'inserimento delle esplicative di interesse ed eventuali loro interazioni, polinomi, ... come suggerito dalle analisi grafiche. Si procede in questo modo se la numerosità del campione lo consente, vale a dire se abbiamo n sufficientemente alto da stimare tutti i coefficienti.
   - In alternativa, si può partire da un modello semplice, con una sola esplicativa, e complicarlo con l'inserimento di altri termini.
   - La modellazione procede con la selezione delle variabili tramite valutazione del p-value del test di significatività del coefficiente associato ad ogni esplicativa, al confronto tra modelli annidati e/o modelli non annidati, alla valutazione della capacità di previsione del modello (grafico valori osservati di Y *vs* valori previsti dal modello), alla valutazione dei residui.
   - Attenzione alle interazioni tra esplicative. L'interazione tra una continua ed una categoriale viene spesso suggerita da un'analisi grafica. Può trattarsi di una interazione significativa o meno, ma rimane tendenzialmente di agevole interpretazione. Lo stesso vale per l'interazione tra variabili categoriali (caso meno frequente): in tal caso, l'unica eventuale complicazione è di calcolo, perchè potrebbero esserci combinazioni di livelli tra le categoriali senza osservazioni (ci si accorge agevolmente del problema perchè `R` non riesce a stimare i coefficienti associati ..... la stima risulta NA). Risulta più difficile invece spiegare l'interazione tra variabili continue. L'interazione può risultare significativa e aumentare l'adattamento del modello...ma come si interpreta? In tal caso ci si domanda se il guadagno in termini di adattamento giustifica l'eventuale minore interpretabilità del modello.

5. Come si comportano i **residui**? Se sono problematici (andamenti deterministici, andamenti non normali, ...) allora il modello va rivisto perchè manca qualcosa....oppure non abbiamo ancora tutte le esplicative a disposizione. Ci sono valori anomali? Se ve ne sono possiamo valutare se la loro eliminazione dal modello crea sostanziali variazioni nelle stime e nelle capacità previsive.

6. Altra possibile strada è la valutazione di qualche **inserimento di splines** (di regressione o smoothing) nel modello. Ad esempio, se il diagramma di dispersione tra Y e X continua indica una relazione non lineare (ad esempio tra `box` e `cmngsoon` nel dataset `film` o tra `box` e `budget` nel dataset `film`) una spline può aggiungere flessibilità al modello e migliorare l'adattamento. Forse (da valutare caso per caso) potrebbe anche risolvere il problema della presenza di residui anomali.

7. Si supponga che la dimensione del problema dipenda da molte esplicative. Possiamo procedere usando una o più tecniche di **riduzione della dimensionalità**

   - selezione automatica: le possibilità forward, backward, mixture dipendono da n e p; le scelte basate su BIC, R^2 aggiustato etc. possono portare a risultati diversi
   - regolarizzazione ridge e lasso

- regressione con le componenti principali: l'interpretazione può non essere agevole

8. **Come scegliamo tra modelli di natura diversa?** Ad esempio, tra un modello lineare ed un modello semiparametrico? Tra una regressione ridge ed una regressione con componenti principali? Le possibilità sono diverse, la valutazione dipende dai modelli a disposizione e da ciò che l'output di un modello fornisce:

   - AIC;
   - valori di validazione incrociata, ad esempio forniti dalle procedure di stime ridge e lasso, o per confrontare modelli non annidati;
   - confronto grafico tra valori di Y e valori previsti dal modello: un buon modello fornisce valori vicini alla bisettrice I-III quadrante;
   - valutazione di MSE su un validation set creato appositamente (approccio più debole di una validazione incrociata ma che fornisce una prima indicazione utile);
   - MSE sui valori del training set $\sum_{i=1}^{n}(valori\ osservati - previsioni)^2/n$ (approccio più debole di una validazione incrociata ma che fornisce una prima indicazione utile).

9. Ricordiamo che i) la selezione automatica tramite le funzioni della libreria `leaps` si applica ai modelli lineari; ii) per la classificazione abbiamo a disposizione anche l'analisi discriminante lineare come metodo alternativo alla regressione logistica (basato su un approccio diverso).

In conclusione, queste note rappresentano esclusivamente una serie di indicazioni sommarie per condurre un'analisi di dati: non rappresentano l'ordine imprescindibile con cui le analisi vanno eseguite, nè un'analisi va necessariamente effettuata applicando tutte le tecniche riassunte. Saranno le caratteristiche e la struttura dai dati, gli scopi dello studio, le relazioni tra le variabili a guidare il processo di costruzione dei modello o dei modelli, nonchè la propensione da parte di chi esegue l'analisi verso una o più tecniche (e la teoria che le sostiene) rispetto ad altre.