Confronto fra modelli di apprendimento supervisionato

- Dati due modelli supervisionati M_1 e M_2 costruiti con lo stesso training set
- Misura della performance di ciascun modello: tasso di errore sul test set
- Confronto fra i modelli come verifica di ipotesi nulla:

non c'è una differenza significativa nella performance dei due modelli M_1 e M_2

Confronto fra modelli di apprendimento supervisionato

- Tre possibili scenari per il test set:
  1. L'accuratezza dei modelli è confrontata tramite due test set indipendenti selezionati casualmente da un insieme di dati campionari
  2. Per confrontare i modelli sono usati i medesimi dati di test; il confronto è effettuato a coppie, osservazione per osservazione
  3. Per confrontare i modelli sono usati i medesimi dati di test; il confronto di basa sulla performance globale di ciascun modello
- Dal punto di vista statistico, l'approccio più corretto è 1. che però è applicabile solo in presenza di un ampia disponibilità di dati di test (reale possibilità di estrarre set indipendenti)
- Nei casi 1. e 3. si può usare una tecnica semplificata

Confronto correttezza globale di due modelli

- Dati:
  S_i = test set utillizzato per valutare il modello M_i (i=1,2)
  n_i = numero di osservazioni nel test set S_i
  E_i = tasso di errore di classificazione del modello M_i

- Calcolo:

$$q = (E_1 + E_2) / 2 \qquad \sigma^2 = q (1 - q)$$

$$P = \frac{|E_1 - E_2|}{\sqrt{\sigma^2 \left(1/n_1 + 1/n_2\right)}}$$

Confronto correttezza globale di due modelli

- Nel caso i test set abbiano la stessa dimensione n (o si usi per entrambi i modelli lo stesso test set):

$$P = \frac{|E_1 - E_2|}{\sqrt{\sigma^2 (2/n)}}$$

- Verifica dell'ipotesi:
  se $P \geq 2$ allora la differenza di prestazioni tra i due modelli è significativa al 95% di confidenza

- Usando test set distinti, si possono scambiare, ripetere i calcoli e usare per il test di significatività la media dei P

Esempio

- Vogliamo confrontare le prestazioni di due modelli M_1 e M_2
- Ciascun test set contiene 100 osservazioni
- L'accuratezza di classificazione è 80% per M_1 e 70% per M_2
- Svolgiamo i calcoli:

$$E_1 = 0.20 \quad E_2 = 0.30$$
$$q = (0.20 + 0.30) / 2 = 0.25$$
$$\sigma^2 = 0.25\ (\ 1 - 0.25\ ) = 0.1875$$

$$P = \frac{|0.20 - 0.30|}{\sqrt{0.1875\ (2/100)}} \approx 1.633$$

- Dato che $P < 2$ la differenza non è considerata significativa

Tecniche di valutazione non supervisionata

- Apprendimento supervisionato
  e clustering (non supervisionato)
  possono essere considerati approcci complementari
  e pertanto utilizzati l'uno per la valutazione dell'altro

- Per la valutazione esterna
  del clustering (non supervisionato)
  possono anche essere utilizzati
  diversi metodi aggiuntivi

Uso del clustering per la valutazione supervisionata

- Le osservazioni usate per l'apprendimento sono offerte alla tecnica di clustering
- Se le osservazioni si raggruppano nelle classi predefinite contenute nel training set, il modello supervisionato è valido
- Possono essere necessarie molte iterazioni per poter fare una valutazione (es. k-means, sensibile alle scelte iniziali)
- La qualità del clustering non garantisce una performance accettabile sul test set, per cui questa tecnica di valutazione è complementare alle altre (es. utile per identificare la ragione del fallimento del metodo supervisionato)
- La validità dell'approccio in genere cala al crescere del numero di classi predefinite nel training set

Valutazione supervisionata del clustering

- L'apprendimento supervisionato può aiutare a spiegare e valutare i risultati del clustering
  - Si designa ogni cluster formato come una classe
  - Si costruisce un modello supervisionato attraverso la scelta di un campione casuale di osservazioni da ciascuna classe (o anche scegliendo le N osservazioni più rappresentative del cluster)
  - Si valuta il modello supervisionato tramite le osservazioni rimanenti

- Dato che le tecniche non supervisionate mancano di uno strumento di spiegazione, la classificazione aiuta a spiegare e analizzare i cluster formati

Altri metodi per la valutazione del clustering

- I metodi per la valutazione interna sono intrinseci alla tecnica di clustering utilizzata (raggiungimento parametri prefissati di ottimalità del risultato)
- I metodi utilizzabili per la valutazione esterna sono indipendenti dalla tecnica di clustering adottata:
  - Utilizzo di un approccio supervisionato, designando tutto il data set come training set (utile se è necessario spiegare le differenze fra i cluster, es. tramite le regole di produzione generate)
  - Definizione di una misura ad hoc di qualità dei cluster (es. da utilizzarsi in algoritmo agglomerativo per valutare il risultati di un metodo tipo k-means, o come punto di partenza per farlo ripartire)
  - Esame dei valori delle variabili dei cluster (se le variabili differiscono in modo significativo, anche le osservazioni lo faranno)