**https://it.milliman.com/it-IT/insight/validazione-efficace-dei-modelli-di-machine-learning**

**Introduzione**

Anche se sono disponibili numerosi algoritmi di Machine Learning, gli stessi possono essere suddivisi in due diverse macro categorie, supervised e unsupervised, che differiscono tra loro nei criteri utilizzati per classificare i dati o nel fornire le predizioni.

Nella prima categoria di metodologie, i supervised, l’algoritmo suddivide i dati in base a una variabile target definibile a priori, che può essere continua o discreta, a seconda dell’algoritmo utilizzato. Per dividere e classificare i dati, l’algoritmo li suddivide in modo iterativo in base a diverse variabili per ridurre al minimo una funzione di costo. Esistono molte funzioni di costo diverse, ma in generale tutte misurano la differenza tra i valori previsti e quelli osservati. Grazie alla loro capacità di generare previsioni accurate, queste tecniche di apprendimento trovano numerose applicazioni nella gestione dei rischi. Alberi di Classificazione e Regressione, Random Forest sono tutte tecniche di apprendimento supervised.

Nella seconda categoria, gli algoritmi unsupervised, non esiste alcuna variabile target definibile a priori in base alla quale i dati sono suddivisi. Al contrario, l’algoritmo raggruppa le osservazioni in base a quanto esse sono simili tra loro rispetto a determinate dimensioni. Queste tecniche vengono spesso utilizzate in campagne di marketing per personalizzare manovre ed interventi pubblicitari per gruppi di consumatori simili. Tecniche di clustering quali la K-means e modelli di Markov sono esempi di tecniche di apprendimento unsupervised.

**Possibili criticità**

Se da un lato questi algoritmi di apprendimento automatico hanno le potenzialità per migliorare i processi aziendali, affinare le previsioni dei risultati e contenere i costi, dall’altro lato presentano potenziali criticità da tenere in considerazione nel processo di validazione del modello previsionale. Di seguito sono riportate tre potenziali criticità comunemente associate a queste tecniche di apprendimento.

Overfitting**(eccessivo adattamento)**

Le tecniche di Machine Learning e gli algoritmi di classificazione sono più sensibili al fenomeno dell’overfitting rispetto ai più tradizionali modelli di previsione. L’overfitting si verifica quando un modello basa le sue previsioni su correlazioni spurie all’interno di un campione di dati anziché su relazioni autentiche esistenti all’interno della popolazione nel suo complesso. Mentre i modelli lineari ed i GLMs possono generare livelli contenuti di overfitting, gli algoritmi di Machine Learning sono ancora più sensibili a questo tipo di criticità, principalmente a causa della mancanza di vincoli parametrici. In assenza di una forma funzionale, questi algoritmi possono utilizzare ogni relazione, lineare e non, tra le variabili nei dati del campione di training, per eseguire raggruppamenti e/o fare previsioni. Il maggior rischio è che ogni singolo campione, su cui viene costruito il modello, abbia le proprie peculiarità che non riflettano puntualmente la vera popolazione, riducendo il potere di replicabilità del fenomeno osservato nella totalità dei dati. Gli alberi di Classificazione e di Regressione sono particolarmente sensibili a questa tipologia di criticità, in quanto possono suddividere i dati fino a raggiungere la classificazione perfetta o quasi perfetta, generando così alcune partizioni fuorvianti.

Se un modello caratterizzato da overfitting viene applicato a nuovi dati appartenenti alla stessa popolazione, può, potenzialmente, produrre previsioni poco accurate. Vista l’importanza che molte aziende attribuiscono nei processi decisionali alle analisi predittive, questa mancata accuratezza può avere effetti molto distorsivi.

**Riduzione della trasparenza**

Le tecniche di Machine Learning riducono, inoltre, la trasparenza del processo interpretativo del modello previsionale. Con le tecniche di regressione più tradizionali, è più intuitivo vedere come interagiscono le variabili all’interno del modello; infatti per valutare la significatività e la direzione di un effetto, è sufficiente esaminare un singolo coefficiente. Se quest’ultimo è positivo, ciò implica una relazione positiva tra la variabile indipendente di interesse e la variabile dipendente, e viceversa. La maggior parte delle tecniche di apprendimento automatico non produce tuttavia risultati così facilmente interpretabili. Alcuni algoritmi, come gli alberi di Classificazione e Regressione semplici, presentano grafici abbastanza comprensibili, ma altri, quali il Gradient Boosting, le Random Forest e le Reti Neurali, funzionano come delle specie di “scatole nere” diminuendo la trasparenza e l’interpretabilità della metodologia sottostante. Sebbene un utente possa inserire i dati e le specificità del modello e pur essendo possibile esaminare i vari steps intermedi generati dall’algoritmo per capire come è in grado di generare le previsioni, spesso è necessario ed opportuno possedere una conoscenza molto approfondita e consolidata di queste tecniche per poter valutare con trasparenza e giudizio critico l’intero processo.

Se non si è in grado di capire in che modo il modello genera i risultati, gli utenti potrebbero ignorare la presenza di distorsioni latenti ed errori, provocati da inadeguati settaggi ad-hoc del modello, i quali comporterebbero un risultato addirittura controproducente perché andrebbero a ridurre le possibilità di aumentare il potere predittivo. Nelle tecniche di regressione tradizionali, uno sguardo alla direzione e alla significatività del singolo coefficiente, agli standard error ed all’adattamento complessivo dei modelli ai dati, consente agli utenti di avere un’idea, per quanto approssimativa, del corretto potere predittivo dei modelli utilizzati. Se la maggioranza dei coefficienti delle variabili risultano non significativi, la bontà di adattamento del modello è inaccurata, o se le variabili non hanno il potere esplicativo previsto, l’utente sa che il modello potrebbe contenere un errore o un settaggio non ottimale. Il modello potrebbe essere stato adattato su dati non adeguati, scarsamente specificati o utilizzati in un contesto non corretto. Con le tecniche di Machine Learning, tuttavia, la mancanza di trasparenza nelle previsioni rende più difficoltosa l’individuazione di questi tipi di criticità.

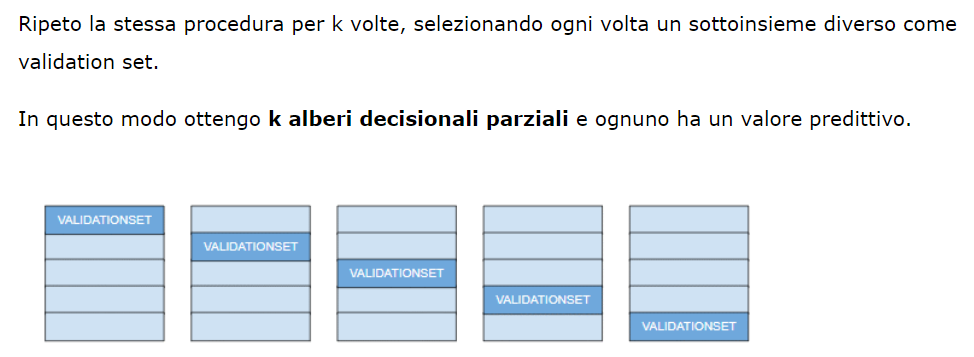
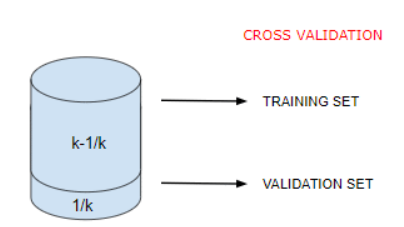
**Risultati che si basano su dati campionari non adeguati**

La qualità dei risultati ottenuti da un modello è direttamente proporzionale a quella dei dati su cui viene costruito. Le tecniche di Machine Learning non fanno eccezione a questa regola. Come detto precedentemente, a causa della scarsa trasparenza di questi algoritmi è facile che agli utenti possano sfuggire potenziali criticità presenti nel processo di modelling, ad esempio che previsioni poco accurate siano dovute da dati campionari non adeguati. Nonostante la gran quantità dei dati disponibili grazie ai progressi conseguiti in campo tecnologico, si devono adottare misure adeguate per assicurarsi che tutti i campioni utilizzati per il modelling siano rappresentativi della popolazione nel suo complesso. Tecniche di campionamento incomplete, o poco idonee al tipo di proiezione che si intende produrre, possono fornire dati non rappresentativi della popolazione nella sua interezza. Oltre alle tecniche di campionamento imperfette o non idonee alla finalità dell’analisi, anche l’utilizzo di dati troppo poco recenti può comportare dei rischi nel processo di modelling. Se un campione utilizzato nel processo di modellazione non rappresenta più il processo sottostante che lo genera, il modello produrrà risultati distorti e non in linea con il trend oggetti di analisi.

**TECNICHE DI VALIDAZIONE DEI MODELLI**

La Cross Validation è una tecnica frequentemente utilizzata per assicurarsi che un modello non produca un eccessivo livello di overfitting rispetto ai dati campionari sui quali è sviluppato. È stata utilizzata in passato per contribuire ad assicurare l’integrità di altri metodi statistici e con la crescente popolarità delle tecniche di Machine Learning, ha visto incrementarsi il suo utilizzo. Con tale tecnica, un modello viene costruito ed adattato utilizzando solo una parte del campione di dati a disposizione, utilizzando la restante porzione per valutarne l’effettivo potere predittivo. In condizioni ideali, il modello funzionerà ugualmente bene su entrambe le porzioni di dati. In caso contrario, è probabile la presenza di un livello di overfitting non trascurabile. I dati campionari vengono solitamente suddivisi in proporzione 80-20, in cui l’80% viene utilizzato per adattare o “addestrare” il modello (training data set) e il restante 20% (validation data set) viene impiegato per valutarne la capacità predittiva. Esistono anche approcci più rigorosi alla Cross Validation, tra cui la k-fold Validation, che prevede che il processo sopra esposto venga ripetuto k volte con diverse suddivisioni dei dati campionari.

Per la valutazione del modello, la Cross Validation può essere utilizzata per confrontare la stabilità del modello previsionale finale e per valutare l’accuratezza delle previsioni per le stesse esposizioni basate sulla modifica del training data set utilizzato. In pratica, si può addestrare ed allenare il modello con differenti tipi di dati campionari, e valutare sequenzialmente le previsioni prodotte al fine di determinare se il modello è eccessivamente sensibile ai dati utilizzati per la fase di training.



A questo punto, calcolo la media dei valori predittivi ottenuti nei k esperimenti. L'albero selezionato è quello che risente meno del rumore causato dagli attributi irrilevanti presenti nel dataset ( overfitting ).