



Data Science and Database Technology

Politecnico di Torino

Data Mining – Esercitazione 4

Obbiettivo

Sfruttare gli algoritmi di classificazione del data mining per analizzare un set di dati reali utilizzando la libreria di machine learning Scikit-Learn su Python.

Per risolvere il laboratorio, è necessario utilizzare il notebook Python disponibile sul [Notebook Google Colab](#) (creare una copia sul proprio account Google prima di modificarlo).

[Guida Google Colab](#)

Dataset

Il dataset Utenti (Users.xls, contenuto nel file zip) raccoglie dati censuari sugli utenti americani di una determinata azienda. Gli utenti sono classificati come “basic” o “premium” in base ai servizi comunemente richiesti. Ogni record del dataset corrisponde a un utente diverso. Il dataset raccoglie circa 32.000 utenti diversi, tra cui alcune informazioni personali dell'utente (ad esempio, età, sesso, classe di lavoro) e la classe corrispondente. L'attributo classe, che sarà utilizzato come attributo di classe in tutto lo studio, è riportato come attributo dell'ultimo record.

L'elenco completo degli attributi del dataset è riportato di seguito.

- (1) Age
- (2) Workclass
- (3) FlnWgt
- (4) Education record
- (5) Education-num
- (6) Marital status
- (7) Occupation
- (8) Relationship
- (9) Race
- (10) Sex
- (11) Capital Gain
- (12) Capital loss
- (13) Hours per week
- (14) Native country

Contesto

Gli analisti vogliono prevedere la classe dei nuovi utenti, in base alle caratteristiche degli utenti già classificati. A questo scopo, gli analisti utilizzano tre diversi algoritmi di classificazione: un albero decisionale (Decision Tree), un classificatore bayesiano (Naïve Bayes) e un classificatore basato sulla distanza (K-NN). Il set di dati degli utenti viene utilizzato per addestrare i classificatori e convalidarne le prestazioni.

Goal

Lo scopo di questa pratica è generare e analizzare diversi modelli di classificazione e convalidare le loro prestazioni sul set di dati Users utilizzando la Scikit-Library su Python. Diversi. Per valutare le prestazioni della classificazione, è necessario testare e confrontare diverse impostazioni di configurazione. Per convalidare le prestazioni dei classificatori è necessario utilizzare un processo di 10-fold Stratified Crossvalidation. I risultati ottenuti da ciascun algoritmo devono essere analizzati per valutare l'impatto dei principali parametri di input.

Domande

1. Imparare un albero decisionale utilizzando l'intero dataset come dati di addestramento e le impostazioni di configurazione predefinite per l'algoritmo Decision Tree. (a) Quale attributo è ritenuto il più discriminante per la previsione della classe? (b) Qual è l'altezza dell'albero decisionale generato? (c) Trovare un esempio di partizione pura nell'albero decisionale generato.
2. Analizzare l'impatto dei parametri di minima impurità (utilizzando il criterio dell'entropia) e di massima profondità sulle caratteristiche del modello di albero decisionale appreso dall'intero set di dati (mantenere la configurazione predefinita per tutti gli altri parametri).
3. Cosa succede se cambiamo l'etichetta della classe da "Classe di servizio" a "Paese nativo"? Rispondete nuovamente alla domanda (1) in questo nuovo scenario.
4. Considerando nuovamente la classe di servizio come attributo della classe ed eseguendo 10-fold Stratified Crossvalidation, qual è l'impatto dei parametri di minima impurità e massima profondità sull'accuratezza media ottenuta dall'albero decisionale? Confrontare le matrici di confusione ottenute utilizzando diverse impostazioni dei parametri: 1) mantenendo la configurazione predefinita per tutti gli altri parametri 2) utilizzando almeno 4 valori diversi (scegliendoli con saggezza!) per ciascun parametro.
5. Considerando il classificatore K-Nearest Neighbor (K-NN) ed eseguendo 10-fold Stratified Crossvalidation, qual è l'impatto del parametro K sulle prestazioni del classificatore? Confrontate le matrici di confusione ottenute utilizzando diversi valori del parametro K. Eseguite 10-fold Stratified Crossvalidation con il classificatore Naïve Bayes. Il K-NN ha un rendimento medio migliore o peggiore rispetto al classificatore Naïve Bayes sui dati analizzati?
6. Analizzare la matrice di correlazione per scoprire le correlazioni a coppie tra gli attributi dei dati. Considerando i risultati ottenuti, l'ipotesi di indipendenza Naïve è valida per il set di dati degli utenti?