Data science e tecnologie delle basi dati

Politecnico di Torino

Homework 2

Obiettivo

Sfruttare gli algoritmi di classificazione del data mining per analizzare un set di dati reali utilizzando Python.

Per risolvere l'homework, è consigliabile utilizzare **Google Colab**. Per comodità, è possibile partire dal template disponibile nel **Laboratorio 4**.

> Guida all'uso di Google Colab

Set di dati

Il dataset Breast (Breast.xlsx, disponibile sul sito web del corso) raccoglie dati medici su pazienti che hanno contratto un tumore al seno. Ogni record del dataset corrisponde a una paziente diversa e consiste in un insieme di caratteristiche della paziente, del trattamento e della malattia (ad esempio, l'età della paziente, le dimensioni del tumore). A seconda che il tumore sia un evento ricorrente o non ricorrente nella vita della paziente, ogni record è anche etichettato con l'etichetta di classe "Recurrence events" o "No-recurrence events". L'attributo di questo dato, che sarà utilizzato come attributo di classe in tutto il lavoro domestico, è riportato come ultimo attributo del record.

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

(1) Age
(2) Menopause
(3) Tumor-size
(4) Inv-nodes
(5) Node-caps
(6) Deg-malig
(7) Breast
(8) Breast-quad
(9) Irradiat
(10) **class (attributo classe)**

Contesto

Gli oncologi vogliono prevedere la proprietà di recidiva o meno dei tumori al seno in base alle caratteristiche della paziente, del tumore e del trattamento. A questo scopo, sfruttano tre diversi algoritmi di classificazione: un albero decisionale (Decision Tree) e un classificatore bayesiano (Naïve Bayes), e un classificatore basato sulla distanza (K-NN). Il dataset Breast viene utilizzato per addestrare i classificatori e per validarne le prestazioni.

Domande

Rispondete alle seguenti domande:

1. Addestrare un albero decisionale dall'intero dataset impostando la soglia di profondità massima a 5, mantenendo la configurazione predefinita per tutti gli altri parametri. (a) Quale attributo è ritenuto il più discriminante per la previsione della classe? (b) Qual è l'altezza dell'albero decisionale generato? (b) Individuare una partizione pura nell'albero decisionale e riportare una schermata che mostri l'esempio individuato.
2. Analizzare l'impatto dei parametri impurità minima (utilizzando il criterio di suddivisione dell'entropia), numero minimo di campioni per ogni foglia e profondità massima sulle caratteristiche del modello di albero decisionale appreso dall'intero set di dati (mantenere la configurazione predefinita per tutti gli altri parametri). Riportare almeno 5 diverse schermate che mostrino gli Alberi decisionali (o porzioni di essi) generati con diverse impostazioni di configurazione.
3. Eseguendo una 10-fold cross-validation stratificata, qual è l'impatto dei parametri impurità minima, numero minimo di campioni da dividere e profondità massima sull'accuratezza media ottenuta dall'albero decisionale? Riportare almeno 5 schermate che mostrino le matrici di confusione ottenute utilizzando diverse impostazioni dei parametri (considerare *almeno* tutte le configurazioni utilizzate per rispondere alla domanda 2). Mantenete la configurazione predefinita per tutti gli altri parametri.
4. Considerando il classificatore K-Nearest Neighbor (K-NN) ed eseguendo una 10 fold CrossValidation stratificata, qual è l'impatto del parametro K sull'accuratezza media del classificatore? Riportate almeno 5 schermate che mostrino le matrici di confusione ottenute utilizzando diversi valori del parametro K. Eseguite una convalida incrociata stratificata a 10 volte con il classificatore Naïve Bayes. K-NN si comporta in media meglio o peggio del classificatore Naïve Bayes sui dati analizzati? Riportare una schermata che mostri la matrice di confusione ottenuta da Naïve Bayes sul set di dati analizzato.
5. Analizzare la matrice di correlazione per scoprire le correlazioni a coppie tra gli attributi dei dati. Riportare una schermata che mostri la matrice di correlazione ottenuta. (a) L'ipotesi di indipendenza Naïve è effettivamente valida per il set di dati Breast? (b) Qual è la coppia di attributi più correlata?

Assegnazione

Scrivete una relazione di 4-5 pagine contenente le risposte alle domande di cui sopra, includendo grafici e visualizzazioni. Analizzare le prestazioni e il comportamento dei modelli per le diverse impostazioni.

NOTA

Per qualsiasi problema, scrivete un'e-mail con i seguenti metadati:

A: simone.monaco@polito.it, alkis.koudounas@polito.it

Oggetto: [DSTBD] Bug Homework 2

Corpo: <Descrizione del problema>.