ESERCITAZIONE SUPERVISED LEARNING

Corso di Data Analytics

18 Novembre 2020

Si consideri il dataset Wine (presente sulla pagina di Virtuale del corso):

1. Usando la tecnica del `DecisionTree (max_depth=6)`, e con test-train split del 33% (`random_state=100`), identificare quale `impurity measure` (Gini vs Entropy) sia più performante di termini di accuratezza di classificazione. Visualizzare e confrontare gli alberi decisionali prodotti usando le due metriche.
2. Usando la tecnica del `RandomForest (n_estimators=20, max_depth=4)`, e con test-train split del 33% (`random_state=100`), identificare le tre variabili più importanti per il processo di classificazione. Ripetere la classificazione usando solo tali variabili.
3. Usando la tecnica dei `DecisionTree`, studiare l'impatto del parametro di pruning α sulle prestazioni del classificatore in termini di accuratezza media. Ripetere gli esperimenti sia sul Training sia sul Test Set, utilizzando un test-train split del 33% ed il metodo delle prove ripetute (30 ripetizioni). Considerare le seguenti configurazioni del parametero α : [0, 0.005, 0.01, 0.02, 0.05, 0.1]. Visualizzare il risultato dell'analisi mediante un grafico (asse X→ fattore α, asse Y→ Accuratezza media, due serie dati per Train e Test).
4. Usando la tecnica del `RandomForest`, studiare l'impatto del parametro di `n_estimators` sulle prestazioni del classificatore in termini di accuratezza media. Ripetere gli esperimenti sia sul Training sia sul Test Set, utilizzando un test-train split del 33% ed il metodo delle prove ripetute (30 ripetizioni). Considerare le seguenti configurazioni del parametro `n_estimators` : [1, 10, 20, 50, 100, 200]. Visualizzare il risultato dell'analisi mediante un grafico (asse X→ fattore `n_estimators` , asse Y→ Accuratezza media, due serie dati per Train e Test).
5. Usando la tecnica del `GradientBoostingClassifier`, studiare l'impatto dei parametri `n_estimators` e `learning_rate` sulle prestazioni del classificatore in termini di accuratezza media utilizzando un test-train split del 33% ed il metodo delle prove ripetute (30 ripetizioni). Considerare le seguenti configurazioni del parametro `n_estimators` : [1, 10, 20, 50, 100, 200], e `learning_rate=[0.01, 0.05, 0.1, 0.2]` Visualizzare il risultato dell'analisi mediante un grafico (asse X→ fattore `n_estimators` , asse Y→ Accuratezza media, quattro serie dati per differenti `learning_rate`).
6. Considerando la configurazione ottimale di `n_estimators` e `learning_rate` determinate nel punto precedente, ed un test-train split del 33% (`random_state=100`), visualizzare l'accuratezza media della previsione ad ogni stage del `GradientBoostingClassifier` (Usare il metodo `staged_predict(X)`).
7. Confrontare le prestazioni dei classificatori `DecisionTree` (scegliendo la configurazione ottimale del parametro α di cui al punto 2), `RandomForest` (scegliendo la configurazione ottimale del parametro `n_estimators` di cui al punto 3), `LogisticRegression`, `GradientBoostingClassifier`, `GaussianNB` e `QDA`, in termini di accuratezza media. Visualizzare il risultato dell'analisi mediante un grafico a barre (asse X→ fattore `n_estimators` , asse Y→ Accuratezza media sul Test).