Universidade do Minho
Departamento de Informática
Mestrado [integrado] em Engenharia Informática

Dados e Aprendizagem Automática
1°/4° Ano, 1° Semestre
Ano letivo 2023/2024

Enunciado Prático n° 6
2 de novembro de 2023

Tema

Máquinas de Vetores de Suporte e Árvores de Decisão: Afinação de Hiperparâmetros com *GridSearchCV* e *Pruning*

Enunciado

Máquinas de vetores de suporte são modelos lineares de aprendizagem supervisionada aplicados a problemas de classificação ou de regressão, onde a técnica cria hiperplanos que separam os casos de estudo em classes. O objetivo é encontrar um hiperplano que tenha a margem máxima, ou seja, a distância máxima entre os pontos de dados de ambas as classes. Este modelo permite resolver problemas lineares e não lineares, produzindo modelos com *accuracy* significativa e com menor esforço computacional.

Árvores de decisão são modelos lineares de aprendizagem supervisionada também aplicados a problemas de classificação ou de regressão. Nesta técnica, os dados são continuamente divididos de acordo com um determinado parâmetro, onde as "folhas" representam as decisões ou resultados estimados.

Grid search é uma técnica de afinação de hiperparâmetros que pode facilitar a construção de um modelo e a avaliação de um modelo para cada combinação de parâmetros de algoritmos por grelha.

Pruning é uma técnica de remoção de folhas e subárvores de uma árvore de decisão aplicada quando estas partes não apresentam significativa contribuição para a precisão e/ou interpretabilidade da árvore, reduzindo-se assim a complexidade da árvore e aumenta-se a sua generalização.

Tarefas

Neste enunciado, utilizaremos o *dataset* de disponibilizado (*incidents.csv*), que contém informações de multi-tipo. Para o desenvolvimento de um modelo de classificação, foi decido aplicar-se os modelos máquina de vetores de suporte e árvores de decisão.

Atendendo ao problema em questão, deverão seguir os seguintes passos:

T1. Carregar o *dataset*, utilizando a função *pandas.read_csv(...)*;

T2. Aplicar métodos para exploração e visualização de dados;

T3. Preparar e organizar os conjuntos de casos de estudo do *dataset* em dados de treino e teste, utilizando a função *sklearn.model_selection.train_test_split (..., test_size = 0.3)*;

T4. Treinar um modelo de máquina de vetores de suporte (*sklearn.svm.SVC*) e um modelo de árvore de decisão como classificador (*sklearn.tree.DecisionTreeClassifier*).

Nota: Definir o X e o y. Atenção ao tipo dos atributos que fazem parte do X;

T5. Obter matrizes de confusão e relatório de classificação dos modelos e efetuar a respetiva análise crítica. Avaliar a *accuracy* do modelo na previsão de '*incidents'*. Avaliar também o modelo usando a métrica *f1_macro.* Que conclusões se podem tirar?

T6. Aplicar a técnica de *gridsearchCV* (*sklearn.model_selection.GridSearchCV*) como forma de procurar o conjunto de hiperparâmetros capaz de otimizar o desempenho da classificação dos modelos de máquina de vetores de suporte (*C* e *gamma*) e de árvore de decisão (*criterion* e *max_depth*). Qual a variação no desempenho do modelo subjacente a estas alterações?

T7. Aplicar a técnica de *pruning* como forma de identificar o conjunto de hiperparâmetros capaz de otimizar o desempenho da classificação do modelo de árvore de decisão (*max_depth* e *ccp_alphas*). Qual a variação no desempenho do modelo subjacente a estas alterações?