TP555 - AI/ML

Lista de Exercícios #8

Aprendizagem em conjunto e florestas aleatórias

1. Digamos que você treinou cinco modelos diferentes com exatamente os mesmos dados de treinamento e todos alcançam 95% de precisão, existe alguma chance de você poder combinar esses modelos para obter melhores resultados? Se sim, como? Se não, por que?
2. Qual é a diferença entre classificadores de votação rígida e suave?
3. É possível acelerar o treinamento de um ***bagging ensemble*** distribuindo-o por vários servidores? E quanto ao ***pasting ensemble*** ou ***floresta aleatória***?
4. Qual é o benefício da avaliação ***out-of-bag***?
5. O que torna as ***árvores-extras*** (***extra-trees***) mais aleatórias do que as ***florestas aleatórias*** comuns? Como essa aleatoriedade extra pode ajudar? As ***árvores-extras*** são mais lentas ou mais rápidas que as ***florestas aleatórias*** comuns?
6. Neste exercício você irá comparar a performance de ***árvores de decisão*** com e sem o uso de ***bagging ensemble*** utilizando o conjunto de dados das luas (*moons dataset*).
    a. Gere um conjunto de dados das luas usando: make_moons(n_samples=500, noise=0.30, random_state=42).
    b. Divida-o em um conjunto de treinamento e um conjunto de testes usando: train_test_split (X, y, test_size=0.2, random_state=42).
    a. Plote os dados do conjunto de treinamento em relação às classes a que pertencem. Ou seja, defina marcadores diferentes para identificar cada um das classes na figura. Por exemplo, use círculos para denotar exemplos que pertencem à classe 0 e quadrados para denotar exemplos que pertencem à classe 1.
    c. Instancie, treine e realize a predição com o conjunto de testes utilizando uma ***árvore de decisão***: DecisionTreeClassifier(random_state=42)
    d. Qual a precisão desta classificação?
    e. Instancie, treine e realize a predição com o conjunto de testes utilizando ***bagging ensemble*** com ***árvores de decisão***: BaggingClassifier(DecisionTreeClassifier(random_state=42), n_estimators=500, max_samples=100, bootstrap=True, n_jobs=-1, random_state=42)
    f. Qual a precisão desta classificação?
    g. Para cada um dos 2 classificadores plote as seguintes informações
        - A matriz de confusão.
        - A fronteira de decisão.
        - A curva ROC.
    h. Analisando-se as figuras da fronteira de decisão dos 2 classificadores, qual deles irá generaliza melhor? Por quê?