Machine Learning

2nd Assignment - Shahid Beheshti University

April 18, 2023

Due date: May 9th

***** You are required to write a detailed report for implementation tasks. *****

1. Is it possible for an SVM classifier to provide a confidence score or probability when making predictions on a particular instance? Explain it.
2. What actions should you take if you have trained an SVM classifier using an RBF kernel but notice that it underfits the training set? Would it be appropriate to increase or decrease the value of γ (gamma) or C, or both?
3. What does it mean for a model to be ϵ-insensitive?
4. What is the difference between hard margin and soft margin SVM? When would you use each one?
5. Is a node's Gini impurity generally lower or greater than its parent's? Is it generally lower/greater, or always lower/greater?
6. Is it a good idea to consider scaling the input features if a Decision Tree underfits the training set?
7. How can you use tree-based models for feature selection?
8. How do you tweak the hyperparameters of the following model in mentioned circumstances:
    - AdaBoost - Underfitting
    - Gradient Boosting - Overfitting
9. What is the difference between homogeneous and heterogeneous ensembles? Which one is more powerful?
10. How ROC and AUC are being used in the evaluation of classification performance?
11. How does the threshold value used in classification affect the model's performance? This value specifies a cut-off for an observation to be classified as either 0 or 1. Can you explain the trade-off between false positive and false negative rates, and how the choice of threshold value impacts precision and recall?

12. What is the difference between one-vs-one and one-vs-all multiclass classification approaches in classifiers? Under what circumstances would you use one over the other?
13. In this part, you are going to work with the Vehicle Insurance Claim Fraud Detection dataset. You will implement multiple classification models using the Scikit-Learn package to predict if a claim application is fraudulent or not, based on about 32 features. You are expected:
    - Perform exploratory data analysis on the dataset.
    - Try to tackle the problem using the following models :
        - Logistic Regression
        - SVM
        - Decision-Trees
        - Random Forest
        - Other classifiers: KNN, Naive Bayes, Ensemble models (Extra Point)
    - Use stratified cross-validation to report your models' performance.
    - Check whether this dataset is imbalanced or not, if yes, try some techniques to overcome this issue. (including over-sampling, under-sampling, weight-based approaches, etc.)
    - Try to boost the performance of the SVM and Random Forest models that you have used in the above section by utilizing various methods (including hyperparameter tuning, different preprocessing methods, feature engineering, etc.). Don't limit yourself only to the aforementioned methods, based on the quality of your work, extra scores may be granted.

14. How can you use SVM for anomaly detection? What are the challenges in using SVM for anomaly detection? **(Extra Point)**
15. Implement a Bagging classifier from scratch. You can use sklearn for the base model. Test your model on the Penguins dataset. **(Extra Point)**
16. How do you handle the class imbalance in Ensemble Learning? Provide some techniques and explain their working. **(Extra Point)**