Biomedical Data Science & AI

Exercise sheet 7

SoSe2021

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Vinay Bharadhwaj, Yasamin Salimi

Due date: June 15th

Questions

Exercise 1 - Elastic Net & Nested Cross-Validation (11 points)

1. Using the *titanic_survival_data.csv* dataset, train a logistic regression model with elastic net penalization to demonstrate the pros and cons of the different data splitting methods and give a short description on what you observe.
    a. Report the accuracy of data splitting with a test size of 0.2 and random state as 1. **(1 point)**
    b. Plot the boxplot for the accuracy of the KFold cross validation with 5 splits. **(1 point)**
    c. Plot the boxplot for the accuracy of the StratifiedKfold cross validation with 5 splits. **(1 point)**
    d. Inform yourself about leave-one-out cross-validation (LOOCV). Implement LOOCV and mention the pros and cons of the method. **(2 point)**
2. Use the nested cross validation to train a logistic regression with elastic net penalization (leukemia_small.csv).
    a. Split the data into training and test samples using an appropriate cross validation method, and in the inner loop carry out hyperparameter optimization. **(2 points)**
    b. Compute the area under the ROC curve (AUC-ROC) and the area under the precision-recall curve (AUC-PR). **(1 point)**
    c. Plot separate boxplots for the 2 performance metrics. **(1 point)**
3. In your own words, explain how each of the following metrics can be used to assess the performance of a model and then calculate each metric using the following confusion matrix. **(2 points)**
    a. Recall
    b. F1
    c. Balanced Accuracy (BAC)
    d. Matthews Correlation Coefficient (MCC)

	Predicted No	**Predicted Yes**
Actual No	250	20
Actual Yes	30	100

Exercise 2 - SVM (4 points)

1) Inform yourself about SVM and briefly explain the working strategy of linear SVM and why maximizing the margin is a good strategy. **(2 points)**
2) Inform yourself about the non-linearity problem for classifiers. Briefly explain how SVM uses kernel trick to overcome this issue. **(2 points)**

Exercise 3 - Random Forest (10 points)

For the following questions, use random_seed = 1 for better reproducibility of your answers.

1. Load the *breast cancer* dataset from sklearn to your Jupyter notebook. Use label encoding to convert your target variable “class” into numerical form. Split the dataset using a 5-fold cross validation **(1 point)**
2. Set up a parameter grid and use grid search with 5-fold cross validation to identify the best hyperparameter values used to fit a random forest classifier. **(2 points)**
3. Use the best hyperparameters from 2) to fit the final model. Predict the classes of the test set and count the number of samples assigned to each class. **(2 points)**
4. Print the importance of each feature in descending order. Identify the top 5 features. **(1 points)**
5. Mention a case when permutation feature importance is favored over impurity-based feature importance. Use permutation importance to print the importances of your features in a descending order. Compare your answer with that of 4). Do you notice any differences? **(2 points)**
6. In your own words, explain the bootstrapping technique and mention how random forest benefits from its application. **(2 points)**