Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfadr

CE5501 | Fall 2023

Teaching Assistants

Romina Zakerian
Amir-Hossein Kashani
Mohammad-Ali Rezaei

Assignment (4)

Outlines. In this assignment, some practical implementation skills which needed in this and other courses of this degree are noticed as well as regression topics. Remember that you may need to re-use your implementations of this assignment; so, it is suggested to code in functional.

Deadline. Please submit your answers before the end of December 17th in courses.aut.ac.ir. Other methods like sending via email or in social networks are not accepted and will not be considered.

Assignment Manual

Delay policy. During the semester, you have extra 7 days for submitting your answers with delay. Mentioned time is for all assignments. After that, for each day of delay you loss 20% points of that assignment. After 4 days you miss all points and any submit doesn`t acceptable. Remember that saving this time doesn`t have any extra point.

Sharing is not caring. Students are free to discuss and share their ideas about problems with others. But sharing source codes, solutions, answers and other results is not allowed and based on university`s rule, both sides will be graded zero.

Problems are waiting you. Some problems are required to be implemented within a programming language and obtain some charts, images, results, etc; then discuss about it. These types of questions are tagged by #Implementation. Some other problems are required to be solved or computed by hand or research about them. These types of questions are tagged by #Theorical. You are not allowed to use programming language or other technical tools to answer theorical problems.

Report is the key. All students' explanations, solutions, results, discuss and answers must be compacted into a single pdf report. A clean and explicit report is expected and may followed by extra pts; so, you may need to write any related detail or experience during the solving problems. Report file should started within a cover page that it includes course and assignment information as well as identical details like name, student number and email address. Second page should be table of contents that indicates student`s answer to each question. Please repeat your name and student number in left side of footer in other pages. Also, you are free to write in Persian or English. If typing is bothering you, so write in a paper and put its picture with acceptable readability quality in report file.

Organize the upload items. Students should upload their implementation source codes as well as results and report. You should upload a single .zip file with the following structure:

ML_04_[std-number].zip
- Report
  - ML_04_[std-number].pdf
  - [other material and results]
- Source codes
  - P[problem-number]_[a-z].py
  - P[problem-number]_[a-z].ipynb
  - …

Python is the power. Students are free to use any programming language like python, matlab, C++ , etc. However it is recommended strongly to use python in jupyter notebook environment; so, you may need to upload your .py or .ipynb sources.

Feel free to contact. If you have any question or suggestion, need guide or any comment be comfortable to ask via email as well as Telegram group.

Problem 1: why and how **(15 pts)**

#Theorical

a) What is the difference between classification and regression when using SVMs?

b) When SVM is not a good approach ?

c) Say you trained an SVM classifier with an RBF kernel. It seems to underfit the training set: should you increase or decrease γ (gamma)?

d) Should you use the primal or the dual form of the SVM problem to train a model on a training set with millions of instances and hundreds of features?

e) Since ensemble learning provides better output most of the time, why do you not use all of the time? In what situations do you not use ensemble classifiers?

Problem 2: Compare Kernels (25 pts)

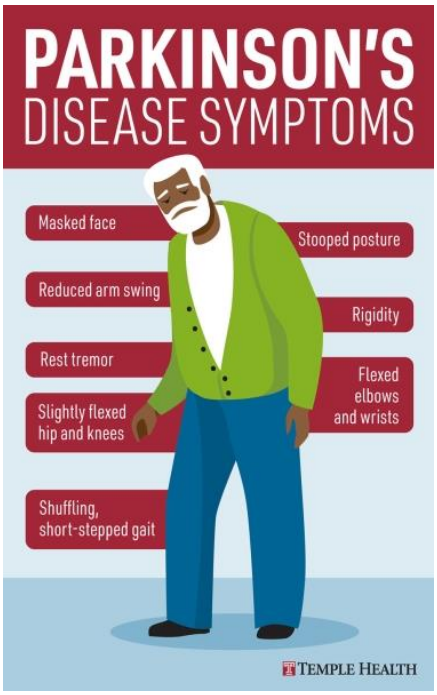
#Implementation

In this question, we are going to compare kernel performance that is mentioned below (use 5-fold cross-validation to evaluate the model):

- Linear kernel
- Polynomial kernel
- RFB
- Sigmoid

Dataset:

The **LSVT Voice Rehabilitation** dataset is a collection of speech recordings of individuals with Parkinson's disease (PD) who have undergone LSVT Voice Rehabilitation (LSVT LOUD) therapy. The dataset includes measurements of various acoustic features of speech, such as fundamental frequency (F0), intensity, and spectral features. It also includes demographic information and clinical assessments of speech quality.



a) Download the data from the link.("Data" sheet includes feature and "Binary response" includes labels)
b) Find appropriate parameters to achieve the best result in accuracy.
c) Find appropriate parameters to achieve the best result in F1-score.
d) Describe the method used to find the best parameters.
e) Report best model.

Problem 3: classifying Spam (20 pts)

#Implementation

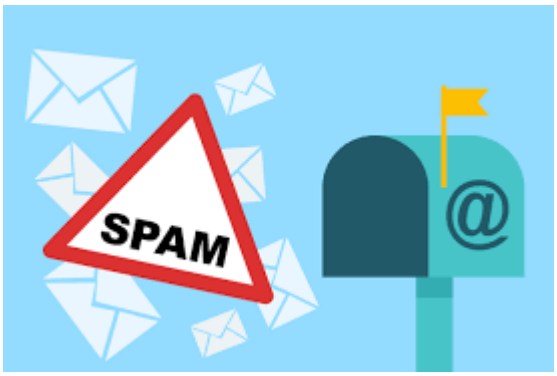


Figure 1: Spam image

In the digital realm, spam, characterized by unsolicited and often malicious messages, poses a persistent threat to email communication.
From unwanted advertisements to phishing schemes, spam takes various forms, making it imperative to detect and classify such content effectively.
In machine learning tasks, this process is crucial for several reasons. Firstly, it enhances user experience by filtering out irrelevant and potentially harmful messages, fostering a safer online environment. Secondly, detecting spam is essential for maintaining the integrity of data used to train machine learning models, ensuring they learn from authentic and representative samples. Moreover, accurately classifying spam aids in preserving the trustworthiness of communication channels, preventing users from falling victim to scams or compromising their security. As we navigate the intricate landscape of digital communication, the integration of robust spam detection mechanisms becomes indispensable for the efficacy and reliability of machine learning applications.

a) Take a look at the spam dataset and shortly describe what kind of classification problem this is. (Python Hint: read spam.csv)

b) Use a decision tree to predict spam. Re-fit the tree using two random subsets of the data (each comprising 60% of observations). How stable are the trees? (Python Hint: Use from sklearn.tree import plot tree to visualize the trees.)

c) Forests come with a built-in estimate of their generalization ability via the out-of-bag (OOB) error.
Use the random forest learner, (Python: RandomForestClassifier()) to fit the model and state the out-of-bag (OOB) error. Explain more about this error.

d) You are interested in which variables have the greatest influence on the prediction quality. Explain how to determine this in a permutation-based approach and compute the importance scores for the spam data.

(Python Hint: choose an adequate importance measure as described in https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html)

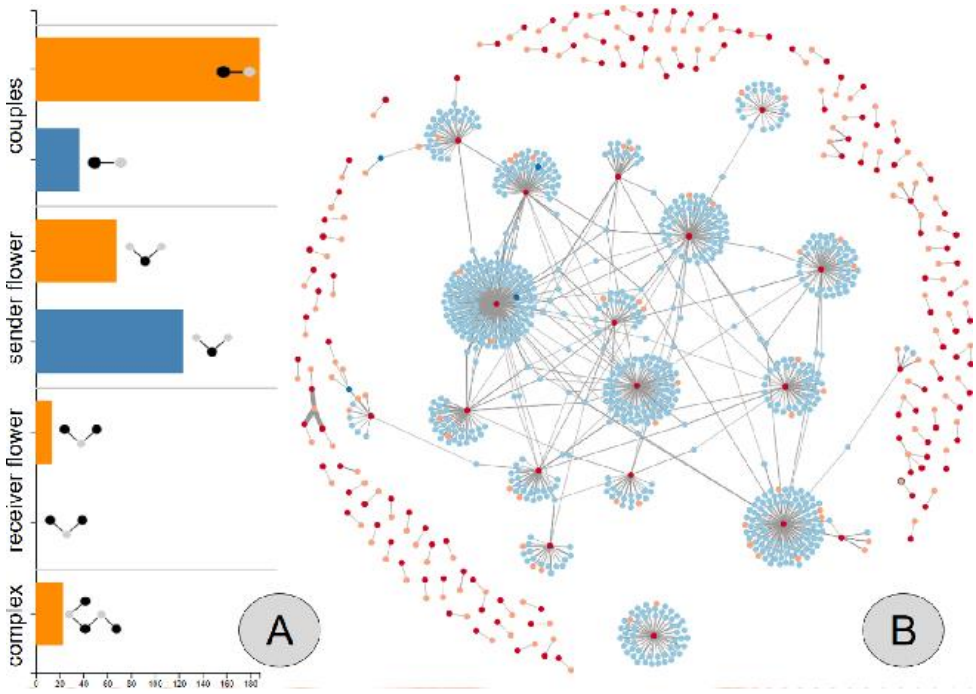
Problem 4: Comparing Ensemble Methods for Fraud Detection (20 pts)

#Implementation

In this question, you will compare the performance of different ensemble methods for fraud detection. Specifically, you will compare the following ensemble methods:

- Random Forest
- Gradient Boosting Trees (XGBoost)
- AdaBoost

The dataset for this task is the Credit Card Fraud Detection dataset, which is publicly available on the link. This dataset contains transactions from a credit card company, with labels indicating whether each transaction is fraudulent or not.



Tasks:

a) Data Preparation: Download the Credit Card Fraud Detection dataset. Preprocess the data as needed. This may include handling missing values, normalizing numerical features, and encoding categorical features (describe your methods in your report completely).

b) Model Training and Evaluation: Train each of the three ensemble methods (Random Forest, XGBoost, and AdaBoost) on the training data. Use 5-fold cross-validation to evaluate the performance of each model on the training data. Select the best parameters for each model based on the evaluation results.

c) Model Comparison: Compare the performance of the three models on the test data using the F1-score and ROC AUC metrics.Discuss the strengths and weaknesses of each model.

d) Hyperparameter Tuning: Describe the method used to find the best parameters for each model. Discuss the challenges of hyperparameter tuning and how you addressed them.

Additional Considerations:

- Consider using grid search or random search to find the optimal hyperparameters for each model.
- Visualize the performance of the models using ROC curves and confusion matrices.
- Discuss the trade-offs between the different ensemble methods.

Problem 5: SVM **(20 pts)**

#implementation

The objective of this project is to predict a binary outcome based on various parameters using a Support Vector Machine (SVM) model. The file of the dataset is in the folder of the exercise with the name “dataset_ml_heart”. The dataset contains 14 columns, each representing a different parameter. The 'T' column represents the binary outcome we want to predict.

Steps:

1. Data Loading and Exploration: Load and understand the dataset.
2. Data Preprocessing: Clean and prepare the data for modeling.
3. Data Visualization: Visualize the data to identify patterns and correlations.
4. Data Splitting: Divide the data into training and test sets.
5. Data Scaling: Standardize the range of the features.
6. Model Training: Train the model on the training data.
7. Model Evaluation: Assess the model's performance using the test data.

a) What preprocessing steps were necessary to prepare the dataset for modeling?

b) What is the purpose of the regularization parameter (C) in the SVM model?

c) What is the accuracy of the model on the test set? How does it compare to the performance on the training set?

d) How can you improve the model's performance?