Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2022

Teaching Assistants

Mohsen Ebadpour(head) (M.Ebadpour@aut.ac.ir)

Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)

Mohamadreza Jafaei (Mr.Jafaei@aut.ac.ir)

Assignment (4)

Outlines. In this assignment, SVMs and Ensemble methods are noticed.

Deadline. Please submit your answers before the end of December 23rd 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 (35 pts)

#Theorical

a) Which method of ensembles can be the better choice for noisy datasets? Explain your reason.

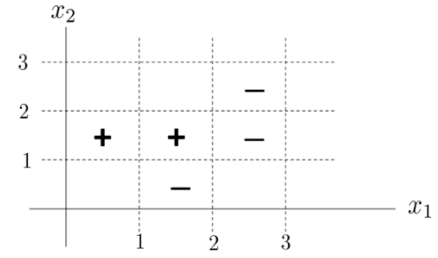
b) In Bagging, what is the difference if we select m random samples with replacement or not?

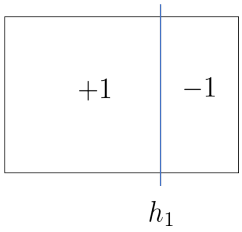
c) The power of ensemble methods is using multiple weak classifiers; is it a good idea to change them with robust classifiers? Why?

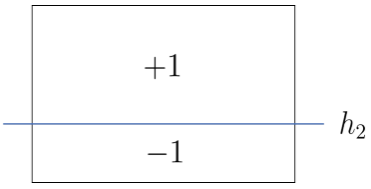
d) Assume that you made a Bagging ensemble for a classifying problem with 50 decision trees with a max deep of four(The problem has seven classes); Also, the dataset consists of 40 features, and the accuracy of each model is less than 14%. Estimate the final accuracy. Is it acceptable? Why? Justify your answer.

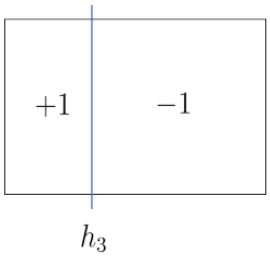
e) Which method of ensembles can be the better choice for the high dimensional dataset with fewer samples and the low dimensional dataset with many samples? Explain your reason.

f) In this part, you are asked to solve a simple Adaboost by hand; You have three weak learners with their decision boundaries respectively; with calculation and indicating all numerical steps, obtain the final classifier. Let the data D consist of five points in the plane as given in below Two points (0.5, 1.5) and (1.5, 1.5) are labeled with +1 mark and three points (1.5, 0.5), (2.5, 1.5) and (2.5, 2.5) are labeled with −1 mark.(h_1 with $x_1 = 2$, h_2 with $x_2 = 1$ and h_3 with $x_1 = 1$)









g) Using Lagrange, find the minimum of the following functions considering its constraints.

1) $f(x, y) = 2 - x^2 - 2y^2$

$g(x, y) = x + y - 1 = 0$

2) $f(x, y) = x^2 + y^2$

$g_1(x, y) = x + 1 = 0$

$g_2(x, y) = y + 1 = 0$

3) $f(x, y) = x^3 + y^3$

$g(x, y) = x^2 - 1 \geq 0$

$g_2(x, y) = y^2 - 1 \geq 0$

h) using SVM, determine the hyperplane separating these two classes; indicate calculation.

(1, 1) => +1

(2, 2) => -1

i) For each of the functions K below, state whether it is a kernel or not. If you think it is, prove it; if you think it is not, give a counter-example.

1) $K(x, z) = K_1(x, z) + K_2(x, z)$

2) $K(x, z) = K_1(x, z) - K_2(x, z)$

3) $K(x, z) = \alpha K_1(x, z)$

4) $K(x, z) = K_1(x, z)K_2(x, z)$

j) One the most important disadvantages of SVM classifiers is that when a single sample added to dataset, calculation of decision boundary and its margin should be repeated(why?). suggest a simple algorithm to avoid this.

k) what is the Karush-Kuhn-Tucker Conditions.

Problem 2: SVM with custom kernel (15 pts)

Implementation

According to the functions in 1.i that you showed as the kernels, build custom kernels and evaluate the performance of them on the iris dataset with read-to-use SVC model. (load it from sklearn and split it as an 8:2 ratio of train and test).

K_1 = RBF Kernel

K_2 = Sigmoid Kernel

Problem 3: SVM outlier detection (15 pts)

Implementation

a) Use the following function to generate a custom dataset:

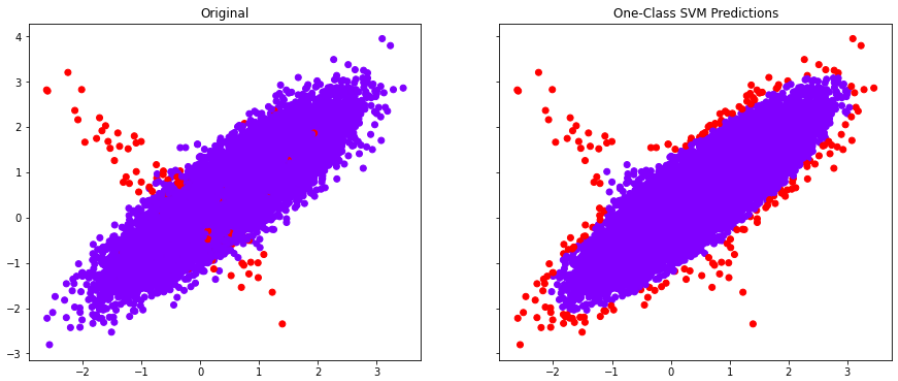
make_classification (n_samples=100000, n_features=2, n_informative=2, n_redundant=0, n_repeated=0, n_classes=2, n_clusters_per_class=1, weights=[0.995, 0.005], class_sep=0.5, random_state=42).

By doing this you will generate 100,000 synthetic data such that 99.5% of the data belongs to class 0 and 0.05% belongs to class 1. In fact, our class 1 is synthetic outliers. split the dataset into 80% training data and 20% validation data.

b) What is one-class SVM and how can it be used in outlier detection? run the one-class SVM on training data. (Set the nu, kernel and gamma).

c) Make outlier predictions on the validation dataset. Which of the metrics such as precision, recall, etc. can show how accurately our model was able to detect outlier data? Explain your reasons.

d) In this section, plot non-outlier and outlier data for validation data. Like the figures below:



Problem 4: Create an anti-malware (35 pts)

Implementation

Malware programming is one of the most adversarial ways to earn money, but it was popular in 2014(based on virustotal.com report). As you guessed, anti-malware programming had the same behavior to make money with. based on two feature sets, can detect malware; one of them is based on static behavior (which accessible before the running it; like file signature or reserved memory allocation) and the other one is dynamic behavior (which accessible after the running it; like network communication or OS`s API call) You are given the custom dataset that consists of OS`s API call behavior; follow the below parts and try to present an ant-malware model. You are allowed to use libraries. (the label is "OUTPUT")

a) Which kind of error is more important in anti-malware applications (both security and user experience sides)? False positive rate or False Negative rate? why?
b) What is stratified splitting, and when should it be used? Using it, drop half the dataset (to reduce the computational cost; you can keep it if you have appropriate resources) and split it to train/test with the ratio of 8:2.
c) Use the Bagging classifier to detect malware. First, use the K-NN for the base estimator, and second, use the decision tree. Try to catch the best result for each one(by fine-tuning the parameters) and compare the results. (Do not forget that set the count of models in ensembles is essential)
d) Next, use the Adaboost to detect malware. First, use the SVC for the base estimator, and second, use the decision tree. Try to catch the best result for each one (by fine-tuning the parameters) and compare the results. (Do not forget that set the count of models in ensembles is essential)
e) Using the replacement and stratified splitting, create 100 datasets, each with 5% of the original dataset. (Create from the training set)
f) Now, train 55 decision trees, 20 SVCs, 15 multinomial naive Bayes classifiers, five K-NNs, and five logistic regressions. For each of mentioned models, use one of the generated datasets. (remember to use weak classifiers)
g) Now, we want you to generate a binary dataset based on the results of weak models and use it to predict the final target (like a weighted Bagging in which weights are obtained by another model). For this aim, pass each sample(both train and test) to 100 above models and store their prediction as a 100-D vector to the related sample. In the end, you should have a table that consists of 101 columns (100 models + label), and its rows are equal to the count of the dataset. (do not combine the test and train sets)
h) In the end, train an SVC and a logistic regression classifier with the obtained binary dataset (use the train set for training and test set for testing). Compare the results with each other and previous parts. which one is better, in your opinion? why?

In each model you trained, report normalized confusion matric with recall, accuracy, and ROC plot.