به نام خدا

دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)

دانشکده برق

یادگیری ماشین – نیم‌سال دوم 1401-1402

تمرین تئوری و عملی ماشین بردار پشتیبان (SVM) ، سری چهارم، درس یادگیری ماشین

Code for all programming assignments should be well documented. **A working program with no comments will receive only partial credit**. Documentation entails writing a description of each function/method, class/structure, as well as **comments throughout the code** to explain the program flow. Programming language for the assignment is **Python**.

Following libraries can be used when necessary:

- Sklearn, Matplotlib, NumPy, SciPy, Pandas and other basic libraries.

Collaboration Policy

You are to complete this assignment individually. However, you may discuss the general algorithms and ideas with classmates, TAs, peer mentors and instructor in order to help you answer the questions. But we require you to:

- not explicitly tell each other the answers
- not to copy answers or code fragments from anyone or anywhere
- not to allow your answers to be copied
- not to get any code from the Web

If you have any questions regarding this assignment, please contact Mr. Ahmadi. (last question: Mr. Janani and Mr. Aghdasian)

Telegram ID: @pouriaA79 @pooya_9877 @A_Aghdasian

Submit by 6th Khordad 1402, 11.59pm

Question 1: SVM for classification [10 point]

Please answer the following questions regarding support vector machines:

a) In the linearly separable case if one of the training samples is removed, will the decision boundary shift toward the point removed or shift away from the point removed or remain the same? Justify your answer. Now if we consider What the decision boundary is of Logistic Regression, will the decision boundary change or remain the same? Explain you answer. (No need to mention the direction of change).

b) Compare Hard SVM and Logistic Regression when the two classes are linearly separable. Give any significant differences. (Hint: think in terms of decision boundary).

c) In the primal optimization of SVM. what's the role of the coefficient C? Briefly explain your answer by considering two extreme cases, i.e., $C \rightarrow 0$ and $C \rightarrow \infty$.

d) Compare Soft SVM and Logistic Regression when the two classes are not linearly separable. Give any significant differences.

Question 2: SVM [10 point]

Suppose we have a binary classification problem with two classes, labeled +1 and -1. We have a training set with three data points:

(1, 2) labeled +1

(2, 3) labeled -1

(3, 4) labeled +1

We want to train a linear SVM to classify new data points. To do this, we need to find the optimal hyperplane that separates the two classes.

a) Compute the weight vector w and bias term b for the optimal hyperplane using the hard-margin SVM formulation.

b) Compute the weight vector w and bias term b for the optimal hyperplane using the soft-margin SVM formulation with C=1.

Question 3: Margins & kernels [10 point]

Please answer the following questions regarding support vector machines:

a) Provide an example scenario for a multi-class SVM, such as with SVM describing m classes.

b) What is meant by a hard-margin SVM model?

c) Suppose you have trained a linear SVM with a boundary and have noticed that it is suffering from underfitting. How would you adjust the model parameters to address this issue?

d) What is a function kernel? Name two common examples and briefly explain the transformation associated with each.

Question 4: SVM [70 point]

In this question, with the MIT-BIH Arrhythmia Dataset, you must implement the SVM algorithm to classify the heartbeat class. The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmias and myocardial infarction. These signals are preprocessed and segmented, with each segment corresponding to a heartbeat.

a) First, randomly select 70% of the data as training data. Then build an SVM learner with linear SVM algorithm and train data set. Calculate the classification accuracy on train and test sets. Then specify the number of support vectors.

b) Build an SVM learner with the nonlinear SVM algorithm in SVM soft mode. Adjust the SVM soft parameter using the validation data you create from the training set to give the best accuracy on the validation set. Plot the curve of changes of classification accuracy on

validation data based on C value. Calculate the accuracy of the classification on the train and test sets in this best selected value. Then specify the number of support vectors and compare with the previous state.

c) Build an SVM learner with the nonlinear SVM algorithm in kernel mode. Choose RBF kernel and polynomial kernel. You can choose the best kernel based on the accuracy of the validation set. Calculate the classification accuracy on train and test sets in this best selected kernel. Then specify the number of support vectors.

d) Build an SVM learner with the non-linear SVM algorithm in the combined mode of kernel and SVM soft. Choose the best kernel from the previous step. Adjust the SVM soft parameter using the validation data to give the best accuracy on the validation set. Calculate the classification accuracy on the train and test sets at this best value. Then specify the number of support vectors.

e) Repeat the previous section with 5-fold-cross validation and compare results.

Attach the written programs along with results and figures requested in each step. Also, make a discussion about results and effect of using non-linear SVM.

What to submit:
• Code
• A short write-up about your implementation with results and your observations from each result.

Good luck 😊