Can use any package for this project

Comp 135 Intro ML: Project 2

Tufts 2018 Spring, Instructor: Liping Liu

Due date: Monday, 03/26

1 Introduction

In this project, you are asked to participate a competition that consists of two classification tasks. For each classification task, you are given a training set, from which you should train a classifier and submit it as part of your submission. We will test the accuracy of your classifier on a holdout test set.

To achieve good performance on the test set (and also good grades for this project), you need to try a few learning algorithms for each task.

For each learning algorithm, you may NOT want to train a classifier with a randomly chosen hyper-parameter. Neither you may want to choose the hyper-parameter by checking the training accuracy, as the model will overfit the data. You may want to run the model selection procedure to decide a good hyper-parameter. Actually, the model selection procedure is required and described as follows.

2 Model selection

The procedure of model selection is to choose a learning algorithm and its hyper-parameter by checking the performance of the learned classifier on a single validation set or through the cross validation procedure.

Requirements: In this project, you are asked to try at least three different learning algorithms. For each algorithm, you need to tune its hyper-parameter(s), that is, you need to choose the hyper-parameter from a set of candidate values based the corresponding cross-validation performances of the classifiers trained with these candidate values. The candidate set for each learning algorithm should contain at least four settings. 3 algorithms, 4 settings for the parameter.of each algorithms. 12 in total.

Here is a guideline of how to decide the candidate set of hyper-parameters. Take logistic regression for example: its hyper-parameter λ can take any positive values. Often the time, you want to test a set of hyper-parameters $\lambda_1 < \lambda_2 < \ldots < \lambda_K$ such that the model overfit the data with the smallest value λ_1 but underfit the data with the largest value λ_K. In this way, you can get a sense that the best value is within the range your candidate values, and your final choice of the hyper-parameter will not deviated the best value radically.

3 Submission requirements

1) Command, clear name of function, be clear of your code

You need to follow submission requirements in this section to guarantee that you submission can be evaluated smoothly.

In your submission, you need to include 1) a zip file containing your code, 2) your classifier file, 3) your nickname in a file, and 4) and your report.

Your submitted code should clearly show your model selection procedure for each learning algorithm. In the documentation of your code, you should make clear how your code runs the model selection procedure for a specific learning algorithm and a possible set of candidate hyper-parameter values.

If you use Python, you should serialize your classifier into a file with Python `pickle`. If you use Matlab, you should save your classifier to a file with the function `saveCompactModel`. The size of the classifier file cannot exceed 6M. We will provide a function `report_accuracy` that checks validity of your classifier. The function is provided in both python and matlab code. Please test your file before you submit it.

As we will generate a leaderboard for the competition, we need your nicknames to show the performance of your classifiers. You may want a cool nickname but others cannot identify you from it.

Your report should clearly states what you have done for the project. First, you need to state the learning algorithms you have used and the meaning of their hyperparameters. Note that, the hyperparameter settings in Python matlab functions might be different from what we have talked in our classes. Taking logistic regression for example, the setting of C parameter has equivalent effect of a special setting of our λ parameter, but C tunes the model in an opposite way as λ. You need to show the candidate set of hyper-parameters for each learning algorithm and the reason of your choices. Second, you need to describe your model selection procedure. Your report should be clear enough for one to repeat your experiment by only reading your report. For each learning algorithm and each classification task, you need to plot cross-validation performances against hyper-parameter settings. Third, you should clearly show the setting for the final classifier you submit. You need to estimate a confidence level that your classifier achieves better performance than the threshold from the validation result. You are encouraged to describe any interesting findings from this project. Try to make the report pleasant to read – not only dry results.

4 Evaluation

In this project, your submission is evaluated based on your code, the final performance of your classifiers, and your report.

- You get 35% of full points if you have correctly done model selection procedure for hyper-parameter tuning for each classification task.

- You get 35% of full points if you have tried three different learning algorithms for each classification task.

- You get 10% of full points if the accuracy of your classifier for the first task is over 0.80 AND that for the second task is over 0.9.

- You get 20% of full points if your report clearly shows what you have done and satisfies the requirements stated in the last section.

If you do not get full points for any grading item, you get partial points according to our estimation of the amount of effort.

5 Classification tasks

Below are Duc and Phong's ironic introductions of two datasets.

5.1 Titanic dataset: Dead or Alive?

It was said that after surviving the sinking of the RMS Titanic, Rose returned to England, her heart broken into pieces. Every night she was haunted by the same thought: ”If only he had decided to share the plank with me...”, and yet her love for Jack was so strong that she believed that somehow Jack had survived the accident. So Rose set out to search for Jack. Unfortunately, she could not find his name under the list of survivors so she decided to ask a machine learning expert to help her build a model that can predict if Jack survived the sinking based on data about the survivors and deceased. Your task is build a model that can predict as accurately as possible whether a person survive the Titanic tragedy.

The dataset given to you will consist of 800 rows, and each row represents a passenger that boarded the RMS Titanic. Each passenger's information is made up of 11 features, which are (starting from left to right):

- Survived: Whether the passenger survived or not, has 2 types 0 = deceased, 1 = survived (This is the label, which is what we want to predict)
- Pclass: Ticket class of the passenger in 3 types: 1 = 1st class, 2 = 2nd class, and 3 = 3rd class.
- Sex: Sex of the passenger in 2 types: 0=Female, and 1=Male.
- Age: Age in years of the dat
- aset.
- sibsp: The number of siblings/ spouses aboard the Titanic
- parch: The number of parents/ children aboard the Titanic
- ticket: Ticket number
- fare: Passenger fare
- cabin: Cabin number
- embarked: Port of Embarkation in 3 types: 0 = Cherbourg, 1 = Queenstown, and 3 = Southampton.

- magic_feature: a magic binary feature, taking value 0 or 1. (The original dataset does not have this feature, but we use a magic function to generate this feature so to make classifier easier).

*Note: there are two things of note in this dataset that you should think about to improve your model accuracy. The first thing is that some of the features might be irrelevant to the training task. The second thing is that each passenger information is represented by a vector of both categorical feature and real-valued features. You should think about how to deal with these problems because they can affect the accuracy of your model.

特征选择+Normalize

6 MNIST dataset - patients' story

It is a proven fact that doctors have bad handwritings, and people have come to terms with it. But Dr. Drake Nguyen's handwriting is something that even the most patient patient in the world will give up trying to comprehend out of frustration. Patients at Tufts Medical Center, tired of deciphering his handwriting, decide to hire you to build a machine learning model to help recognize Dr. Drake Nguyen's handwritten digits based on a dataset provided by the doctor himself. If you succeed, the number of people who overdose on medications because they misinterpret his handwriting will greatly reduce.

The make the problem a binary classification problem, we only classify whether the number being "8" or "9". The dataset given to you is a csv file. Each row is a hand-written number. The first entry is the label, "8" or "9". The rest 784 entries are pixel values of the 28 × 28 image of the written number, and they form the feature vector for the instance.