Knowledge • 0 teams

COL-774, Spring 2017

Competition Details » Get the Data » Make a submission

Assignment of Machine Learning course (COL-774) at IIT Delhi, Spring 2017.

This is the question 3 of Assignment 4 for the Machine Learning course (COL-774) at IIT-Delhi. The goal of this assignment is to experiment with various learning algorithms on a real world dataset.

Dataset Description

You are given a binary classification dataset, and your task is: given features extracted from audio-visual broadcast of news channels, predict whether the broadcast is a commercial (+1) or non-commercial (-1). Each training example is represented as

```
Label <index_1>:<val_1> <index_2>:<val_2> ...
```

where **<index_i>:<val_i>** means **index_i**th feature of the example has value **val_i**. In the dataset, for each example, only features with non-zero values are mentioned in this manner and all other feature values which are not mentioned are assumed to be **zero**. More information on features is available in **attributes.txt** file in data section.

The dataset has been split into training and test sets in the following manner.

-
  ```
  Training  | 77,811 examples with labels
  ```

-
  ```
  Test      | 1,49,137 examples without labels
  ```

More details can be found in the Data section.

Assignment Details

1. [10 points] Train and evaluate using the following algorithms.

(a) SVM (Linear, Gaussian)

(b) Decision Trees & Random Forests

(c) Naive Bayes

Perform 5-fold cross validation on training data to evaluate above models. Depending on the model, try to tune the various parameters, rather than just running a default out-of-the-box version of the algorithm. You are free to use standard implementations of learners in Python & MATLAB (you don't need to implement them on your own though you are free to do so). Report various parameter settings that you try for each of the algorithms and report your cross-validation accuracies.

Please make sure that (a) the libraries are installed on your laptop, so that we can ask you train them during the demo if needed; (b) you have separate scripts for training the model and evaluating the model on the validation sets that we can run directly from the terminal.

2. [15 points] Try out algorithms of your choice to maximize the accuracy on the test data. Improve your model as much as you can - you are allowed to make use of the training data ONLY to train models. A few things that you can try include: different learning algorithms (whether covered in class or not), feature selection/dimensionality reduction methods (PCA, forward/backward selection) and ensemble methods (bagging and boosting). You are strongly encouraged to make use of online resources to read about some of these additional topics not directly covered in class.

You will submit the labels for the test dataset here (on Kaggle), and your performance (on classification accuracy) will be ranked on the leaderboard. There are 2 leaderboards; a Public leaderboard on which you can monitor your accuracy on 40% of the test dataset; a Private leaderboard (which you cannot see) which will be used to determine the final rankings on the other 60% of the test data.

We will award points for both effort and your final position on the Private leaderboard. It is essential that you include details of anything that you try **in your report** -- *we will only rely on your report to determine what you tried*.

Ends: 11:59 pm, Thursday 27 April 2017 UTC
Points: this competition does not award ranking points
Tiers: this competition does not count towards tiers