Mini Project-2

- Deadline: November 5, 2018

Data

In this mini-project you are expected to perform classification task on CIFAR-10 dataset and report the f1-score as well as the accuracy on the test dataset.

Representation and Dimensionality Reduction

Take at least three representations.

1. Raw Data
2. PCA
3. LDA
4. Any other nonlinear or other embedding

Classifiers

Take four out of five classifiers.

1. CART/Decision Tree
2. Soft-margin Linear SVM
3. Kernel SVM with RBF Kernel
4. MLP
5. Logistic Regression

Use implementations from standard libraries.

Experiments:

1. Split the training data into train, val (80:20).
2. Show overfitting and discuss strategies that can minimize overfitting.
3. Report the accuracy and F1-score on test dataset.
4. Clearly show the analysis/experiments for the hyperparameter selection.

Deliverables

- A technical report with graphs/tables and detailed technical discussions. Sample Table

Classifier	Features	Accuracy	F1-score
Linear-SVM	raw-pixels		
Linear-SVM	principle componets		
MLP			

- code, and associated scripts to reproduce the results. (make sure your scripts will help in hyperparameter search, early stopping etc.)
- A brief discussion (~ 1 page) or summary and practical issues in building a classifier from the data.

PS: The objective of this project is you learn how to play around with several hyperparameters by treating models as a black box and compare how the performance varies.

The focus during marking will be methodology and analysis, not on how accurate your solutions are. In addition to the minimum required above, we encourage you to take aspects of the above problem (hyperparameters, choice of models) etc and vary them to study how it affects the training. These will be eligible for bonus marks.

You may use external libraries like libSVM, SVMLite, PyTorch, Tensorflow, scikit-learn, etc

Note: A starter code is provided. You can build on top of it or write your own code from scratch. Few things to keep in mind when you write the code.

1. Avoid hard coding the hyper parameters. Instead use arparse or fire module to supply the parameters via the CLI. Documentation for both the module are available.
2. Try to come up with your own method and techniques to obtain the best accuracy. Click on the link to view the details of results obtained so far, on the CIFAR-10 dataset. Here are a bunch of examples that you can try.

```
1. _Is there any advantage to using an MLP as a feature extractor and
   using an SVM for the final classification, instead of using linear
   classification by the last linear layer with softmax?_

2. _Vary kernels, which one gets you better results? Why?_

3. _Vary the number of pricipal componets._
```

3. Please avoid plagirism. As mentioned above you can use external libraries. The focus is to learn how to use them, understand their documentation, vary the parameters etc. Coding required in this project is minimal. Try to be orignal and come up with novel solutions.

4. If you are using the starter code then you need to have the following:

    1. python 3.5 or above
    2. scikit learn
    3. tqdm

    You can do a simple pip install to download sklearn and tqdm.