SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

51.504 Machine Learning, Fall 2018

Assignment 1

Last update: Saturday 22nd September, 2018 08:34

Grading Policy and Due Date

- You are required to submit: 1) a report that summarizes your experimental results and findings, based on each of the following question asked; 2) your implementation (source code) of the algorithms.
- You are free to choose any programming language you prefer.
- Submit your assignment report and code to eDimension.
- This assignment is an individual assignment. Discussions amongst yourselves are allowed and encouraged, but you should write your own code and report. Write down the names who you have talked to when doing the homework.
- Submit your assignment to eDimension by 7 October 2018 11.59pm. This is a hard deadline. Late submissions will be heavily penalized (20% deduction per day).

Task: Image Classification

In classes we have discussed the issue of classification as a class of *supervised learning* algorithms. One important application of classification is to identify the underlying topics associated with each image by doing *image classification*.

Unzip `data.zip` in the `hw1` folder. In the `data` folder, there are two main directories: `train` and `test`. They contain data used for training and testing respectively. Each directory consists of four sub-folders, where each sub-folder contains several images related to a particular topic (the four topics are: `airplane`, `automobile`, `bird`, `cat`). The goal of image classification is to learn a model (classifier) based on the labeled images in the `train` folder only, such that the model can be used to predict the correct label (i.e., image class) associated with each new image appeared in the `test` folder.

1. (1 pt) Discuss how to formulate this task as a classification problem – describe what are the inputs (x) and what are the outputs (y). Discuss how you would convert the input image into a feature vector? (For example, you could directly use the concatenated image pixels or you could use the HSV/RGB histogram. Please search on web for possible features you could use.)

2. (1 pt) Consider only the images that appeared in these two sub-folders: `bird` and `cat` (i.e., ignore images from `airplane` and `automobile` for now). Implement the logistic regression/loss algorithm using stochastic gradient descent to perform binary classification based on the data from these two sub-folders. Evaluate the model's performance on the training and test set respectively. Discuss

clearly in your report when to stop the algorithm and whether this has any effect on the performance on the test set.

3. (1 pt) Implement the stochastic gradient descent algorithm that minimizes the loss involving the hinge loss to tackle the same binary classification problem. Evaluate your model's performance using the test data. Discuss its learning behavior: the effect of using different learning rates, how fast the algorithm converges. Discuss when to stop the algorithm and whether this has any effect on the performance on the test set. Also compare its performance with that of the logistic loss.

4. (1 pt) Now, consider the multi-class classification problem (i.e., consider all images from all 4 classes). Use the (k-)nearest neighbor classifier for this problem. Try different values of k and report the classification accuracy/error.

5. (1 pt) Think of another way to perform such a multi-class classification task based on what you have learned. Describe and implement your algorithm, and report its performance on training and test set. Compare its performance with the (k-)nearest neighbor classifier. *Hint 1: How to cast a multi-class classification problem into binary classification problems?*

6. (1 pt) In our class, we only talked about directly using pixels to represent an image. However, this method is not very effective as discussed. It is only one of the many ways to represent images as vectors. Think of some other ways to represent the input images as vectors, and discuss the effect of using such alternative representations when SGD with the hinge loss is used. Try to explain why the performance on the test set becomes better or worse with your alternative representations. (Hint: you could use color histograms like HSV or RGB.) If you have a good feature, you could update your answers in items 2-4.

7. (1 pt) The size of the training set usually has a strong influence on the model performance. Try to download some images from Google Image Search using the four class labels. Then, try to train an image classifier using the augmented training set. The test set should NOT be changed. Test if your classification accuracy is increased or decreased. Why? (Hint: does the web data has a similar data distribution with the previous training set?)