M2015: Machine Learning: Assignment 2

IIIT-Hyderabad
`Instructor:Prof.C.V.Jawahar`

1 Problem Setting

We are well aware of the classification problem in machine learning. We have seen a variety of algorithms that help with representation as well as classification. Let us explore the classification problem in this assignment. Specifically, we will be tweaking the feature representation as well as the classifier to see their impact on performance. We will use MNIST and USPS data sets for this purpose. These two popular data sets can be accessed via the links [1, 2]. The datasets have a pre-defined train and test split. It is expected that you follow this train-test split while performing all your experiments.

2 Problem Definition

2.1 Stage 1

Consider the raw data samples from both the datasets (of original dimension). Find the best classification performance among (i) Nearest neighbour (ii) Linear SVM (ii) RBF (iv) Chi-square and (v) Polynomial kernels.

- Compute the confusion matrix for each classifier. Compare and contrast the results.
- Demonstrate via experiments how cross validation affects the hyper parameters.
- While using Linear SVM perform your experiments using both liblinear [6] and libSVM [7]. Compare the training and testing times for both the methods
- Is libSVM with linear kernel same as liblinear?
- 'The performance of RBF kernel can be tuned by varying values of regularization constant C and a kernel hyperparameter γ'. Observe and report the effect of change in hyper parameters while using non-linear kernels.

2.2 Stage 2

[3] demonstrate the use of improved features and achieve the state of the art performance by using an additive-kernel SVM. Use the code [1] provided and reproduce the results. You may use the raw data as the representation. Out of the many dimensionality reduction and feature representation techniques that you have learnt, can you try three alternate representations and see why they are superior or inferior to [3]?

[1] https://people.cs.umass.edu/ smaji/projects/digits/index.html

- Identify the specific instances/classes where the misclassification increases or decreases. Can you explain the reasoning behind this phenomenon?
- Compare and contrast the performance of the three different representations you have chosen with the raw features.

2.3 Stage 3 [5]

In this problem we see how an imaging system may construct models for handwritten numerals. We will use 3000 images from the MNIST database, each of size 28 × 28 pixels.

You will now consider whether these images lie along some lower dimensional manifold in the 784-dimensional image space. The objective of this exercise is to learn the importance of a distance metric; we shall be considering both the plain euclidean distance, and the tangent distance metric.

Construct the 2-D isomap model for the first 3000 examples in the dataset using euclidean distance. Plot the clusters for a sample of the following digits:

- 1 and 7
- 4 and 9
- All the digits
- Show some of the digits on the maps, as done in [4]

Construct the 2-D isomap model using tangent-distance. Plot clusters for the given digits or set of digits as mentioned above.

2.4 Stage 4

Consider your favourite feature representation (say raw pixels of 784 dimensions or a PCA on top of it). Study the following and explain what you see:

- Linear and nonlinear SVMs. What are the parameters that affect the computational requirements? (memory/storage, time)
- Consider a training scheme:
  (i) Train with only 10% of the training data and test on the rest of the training data.
  (ii) If there are any misclassification, add them and retrain.
  Can this training scheme provide the same performance as training with the full dataset? What are the advantages of scaling like this? What are the disadvantages of training like this. Explain with empirical results

References

1. MNIST Database. `http://yann.lecun.com/exdb/mnist/`
2. USPS Database. `http://www.gaussianprocess.org/gpml/data/`

3. Subhransu Maji, Jitendra Malik: Fast and Accurate Digit Classification - Technical Report. (2009)
4. Joshua B. Tenenbaum, Vin de Silva and J.C. Langford: A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science, 2001
5. Generative Discriminative Learning. http://cse.iitk.ac.in/users/cs365/2013/hw1.html(IIT Kanpur, CS365: Artificial Intelligence)
6. libLinear. https://www.csie.ntu.edu.tw/~cjlin/liblinear/
7. libSVM. https://www.csie.ntu.edu.tw/~cjlin/libsvm/