EEL709: Assignment 2

Weightage: 10%

February 26, 2015

1 Objective

To experiment with the use of SVMs for a multiclass classification problem, and understand the effects of varying various parameters therein.

2 Data

We provide a personalised input file that contains 2500 labeled data points, with 15 features each. You should download your file from `http://web.iitd.ac.in/~sumeet/A2/<EntryID>.mat` (for example, `http://web.iitd.ac.in/~sumeet/A2/2010EE50541.mat`). The file format is compatible with both MATLAB and Python; in case you have any difficulties with it and need the data in a different format, let us know. This file will contain two objects: `data`, which is a 15×2500 matrix, with each row corresponding to a feature and each column corresponding to a data point; and `label`, which is a 1×2500 vector, giving the class label for each data point (there are 10 classes, denoted by the labels 0 to 9).

Each data point is actually a low-dimensional representation of an image. To see a visual reconstruction of the image from this representation, you may use the provided script `http://web.iitd.ac.in/~sumeet/A2/visualizeData.m`.

3 Methodology

Your task is to try and learn an SVM classifier for these images, using just the given features, and thereby also to assess the usefulness of the different features. Here is how you should proceed:

1. Familiarise yourself with a standard SVM library. The recommended one is LIBSVM (`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`), which is available for MATLAB, Python, and many other languages. Figure out how you can set various parameters, such as the value of C for the soft-margin SVM, the choice of kernel function, and the kernel parameters (if any). You may wish to play with a simple toy data set to get a feel for using the library, before you move on to the actual data for this assignment.

2. *Binary classification:* Choose just 2 out of the 10 classes in your data, and train an SVM. Leave some data aside for validation, or ideally, use cross-validation. Study the effects of changing the different parameter values, including the type of kernel function being used. How do they affect the accuracy? Can you distinguish cases of overfitting, underfitting, and good fitting? Also try using only the first 10 features, instead of all 15, and compare the results in the two cases. Repeat this exercise for at least two more pairs of classes out of the 10 given to you. Do you consistently get the best results for the same parameter settings, or does it vary a lot depending on which pair of classes you're looking at?

3. *Multiclass classification:* Now we would like to train a classifier for all 10 classes. Figure out what method(s) your chosen library uses for multiclass classification. For instance, LIBSVM uses *one-versus-one* (see Bishop p. 183). Based on your choice, build a classifier for all the classes, and evaluate

using validation or cross-validation. Again, study the effects of changing the various parameters and kernel function. Try and finetune them to obtain the best possible performance. How do these tuned values compare to what you obtained in the binary classification setting? If there are major differences, what might be the reason? Also, try training the multiclass classifier using only the first 10 features, instead of all 15. How does this affect your results? What does this tell you about the usefulness of the different features?

4. *For extra credit.* There are several things that could be done for extra credit:

   - You could try the above multiclass classification using a different approach. For instance, if you have used LIBSVM's one-versus-one, you can now try *one-versus-the-rest* (Bishop p. 182), and compare the results.
   - You could try further reducing the set of features from 10, and systematically looking at the effect of removing or including different features. Which are the 'best' features, and how can you find them?
   - You could try out any other ideas you have for improving the SVM classification accuracy, or for being able to better assess the contribution of the different features and select the most useful ones.

   You can earn a maximum of 3 extra marks, in addition to the 10 allocated for this assignment, depending on the quantum of extra stuff done and the understanding obtained from it.

4 Evaluation

- You should prepare a report, compiling all your results and your interpretation of them, along with your overall conclusions. In particular, you should attempt to answer all of the questions posed in the previous section. Any graphs or other visualisations should also be included therein. If you wish, you may also include code or other materials which are relevant, though this is not required. The submission link is `http://web.iitd.ac.in/~sumeet/submit.html`; put everything into a single zip file or tarball, and name it as per the instructions given there. The submission deadline is **March 15th, 23:59**. Any late submissions will be penalised.

- The schedule for demos/vivas will be announced by your respective TAs, in advance. If for any reason you cannot attend in your scheduled slot, you must arrange for an alternative slot with your TA well in advance. Last-minute requests for rescheduling will normally not be accepted.