Problem Set 4
Assigned: 04/14/2005
Due: 04/28/2005

Please submit an electronic copy of your writeup and code for each problem to `6869-submit@csail.mit.edu`.

Problem 1 *Support Vector Machine*

In this problem you will use an SVM to classify two types of images: natural landscape and man-made structure. The feature used is a simple global statistic: the histogram of gradient direction weighted by gradient magnitude. (You can use `extractfeature.m` to extract the features from images. 16-bins is found to perform best.) For convenience, features are already extracted from all images and saved in `features.mat`. In this file, `nttrain mmtrain nttest mmtest` are the natural landscape training set, man-made structure training set, natural landscape test set and man-made structure test set respectively. They are all 16×100 matrices, with columns as features. The first columns of `nttrain mmtrain nttest mmtest` correspond to 'train_nt001.jpg', 'train_mm001.jpg', 'test_nt001.jpg' and 'test_mm001.jpg' respectively, and so on.

A widely used SVM implementation is available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. This LibSVM is a C implementation. As linked from the webpage, a Matlab interface package is available at `http://www.ece.osu.edu/~maj/osu_svm`. You will use this package for this problem.

(a) First, do PCA on all features from the 200 training samples together. Project features to the first two principal components. Remember to subtract the mean before projection. The resulting 2D coefficients will be inputs to the SVM. This gives classification in a low-dimensional subspace.

Try both polynomial-kernel and radial basis kernel SVM. For consistency, use C-SVC type of SVM, use +1 as one class label and -1 for the other. For polynomial-kernel SVM, use 3rd-degree kernel with 'Gamma'=1 and 'Coefficient'=0. For radial basis kernel SVM, use

'Gamma'=1. In both cases, You should try to train and test with several different parameter of 'C', such as `[1 10 100 1000]`.('C' sets the trade-off between the size of margin and the error tolerance on training samples.) For each 'C', use the provided function `svm_plot.m`{*preferred*} or `SVMPlot.m` in the package to plot the data, the support vectors and decision boundary. Attach the plots in your write-up.

Write down the error rates on the training set and test set for each 'C'. For each SVM, which 'C' gives the best performance on the test set? Does it give the lowest error rate on the training set? {*In practice, to select reasonable model parameters, techniques such as MDL and cross-validation will be used, but you don't need to implement these here.*}

(b) Now let's examine the role of the parameter 'Gamma' for radial basis kernel SVM. Fix 'C' as 1, try 'Gamma' = `[0.001 1000]`. For each case, draw the decision boundary and write down classification error rates on training and test sets. What do you observe?

(c) The values of the discriminant function indicate the "easiness" of correctly classifying corresponding objects. (The discriminant function is the function inside the `sgn()` of the SVM decision function.) For the positive class, the larger the value is, the easier the sample to be classified. Vice versa for the negative class. Use the provided `svm_discrim_func.m` to compute these values.

For radial basis kernel SVM, identify the hardest samples (one from the positive class, one from the negative class) in the training set and test set respectively. Give a brief justification.

(d) *Extra credit*

Can you think of a representation for the images that might give a better classification performance? Does it?

Problem 2 *Image Segmentation*

In this problem you will compare the performance of two different clustering algorithms for image segmentation: k-means and mean-shift. The k-means algorithm is discussed in Chapter 14 of Forsyth and Ponce. The mean-shift image segmentation algorithm is discussed in the paper, D. Comaniciu and P. Meer: "Robust Analysis of Feature Spaces: Color Image Segmentation",

provided as additional reading.

This problem is structured into two parts. The first part focuses on the implementation of the k-means and mean-shift algorithms as general nD-data point clustering algorithms. The second part applies these algorithms to perform image segmentation.

k-Means and Mean-Shift Clustering

(a) Implement the k-means clustering algorithm described as Algorithm 14.5 in Forsyth and Ponce. Your function should have the following syntax:

$$\texttt{function}\ [\texttt{labels}, \texttt{means}] = \texttt{kmeans}(\texttt{data}, \texttt{k})$$

where `data` is a matrix storing the n-dimensional data as its column vectors and `k` is the number of desired clusters you wish the algorithm to form. The cluster labels are returned as `labels`, a vector that has an entry for each data column of `data` storing for each data point an associated cluster label. `means` is a matrix storing the centers of each cluster as its columns.

The MATLAB data file `pts.mat` stores a set of 3D points belonging to two 3D Gaussians. Test and debug your algorithm using this data set with $k = 2$. Plot the resulting clusters using the provided function `plot3dclusters`.

(b) The mean-shift algorithm is discussed in the paper by Comaniciu and Meer assigned as additional reading. Below we provide the details of an implementation of the mean-shift algorithm.

As discussed in the paper, the mean-shift algorithm clusters an n-dimensional data set by associating each point to a peak of the data set's probability density. For each point, mean-shift computes its associated peak by first defining a spherical window at the data point of radius r and computing the mean of the points that lie within the window. The algorithm then shifts the window to the mean and repeats until convergence, i.e. the shift is under some threshold ϵ (we found $\epsilon = 0.01$ to work well). With each iteration the window will shift to a more densely populated portion of the data set until a peak is reached,

where the data is equally distributed in the window. Implement the peak searching processes as the function

$$\texttt{function peak} = \texttt{findpeak}(\texttt{data}, \texttt{idx}, \texttt{r})$$

where `data` is the n-dimensional data set as before, `idx` is the column index of the data point for which we wish to compute its associated density peak and `r` is the search window radius. The algorithm's dependence on r will become apparent from the experiments performed below.

Implement the mean-shift function, which calls `findpeak` for each point and then assign a label to each point according to its peak. This function should have the syntax

$$\texttt{function } [\texttt{labels}, \texttt{peaks}] = \texttt{meanshift}(\texttt{data}, \texttt{r})$$

where `labels` are the same as for part (a) and `peaks` is a matrix storing the density peaks found using meanshift as its columns. Note the mean-shift algorithm requires that peaks are compared after each call to `findpeak` and for similar peaks to be merged. For our implementation of meanshift, we will consider two peaks to be the same if the distance between them is $\leq r/2$. Also, if the peak of a data point is found to already exist in `peaks` then for simplicity its computed peak is discarded and it is given the label of the associated peak in `peaks`.

Debug your algorithm using the data set from part (a) with $r = 2$ (this should give two clusters). Plot your result using the `plot3dclusters` function.

(c) Using the data set stored by `pts.mat` run k-means with $k = 1, 2, 4, 8$ and mean-shift with $r = 1, 2, 4, 5, 10$. Plot the results of k-means and mean-shift in two separate subplots labelling each plot with its corresponding k or r value. Observe the results and discuss the differences between each of the algorithms. How is varying the search window radius r in mean-shift different from varying the k parameter of k-means? How are they the same?

(d) As implemented, the mean-shift algorithm of part (b) is too slow to be realized for image segmentation. We will therefore incorporate the following two speedups into our implementation. Upon finding a peak,

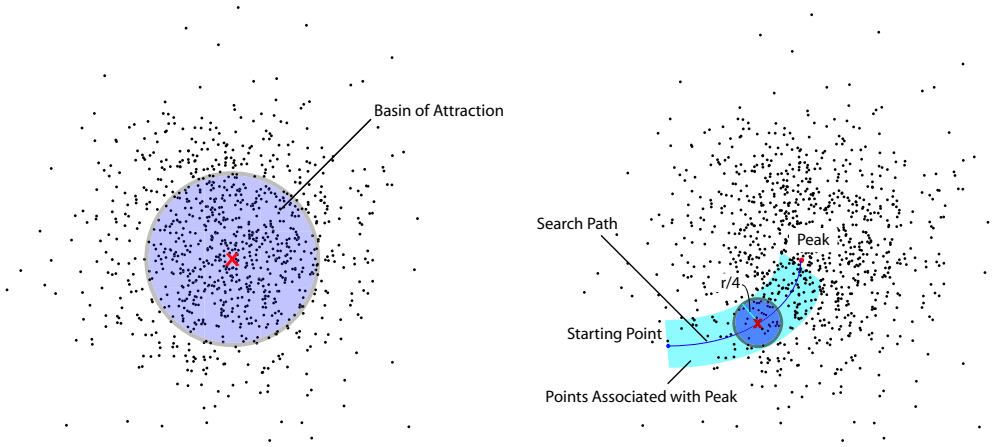

Figure 1: Speedups incorporated into the mean-shift algorithm: (a) basin of attraction, (b) points along the search path are associated with the converged peak.

the first speedup will be to associate each data point that is at a distance $\leq r$ from the peak with the cluster defined by that peak. This speedup is known as *basin of attraction* and is based on the intuition that points that are within one window size distance from the peak will with high probability converge to that peak (see Figure 1(a)). The second speedup is based on a similar principle, where points that are within a distance of r/c of the search path are associated with the converged peak, where c is some constant value (see Figure 1(b)). We will choose $c = 4$ for this problem.

Incorporate the above speedups into your mean-shift implementation by modifying your implementation from part (b). The resulting modified function should have syntax

$$\texttt{function}\ [\texttt{labels}, \texttt{peaks}] = \texttt{meanshift\_opt}(\texttt{data}, \texttt{r}).$$

To realize the second speedup you will need to modify `findpeak` as follows:

$$\texttt{function}\ [\texttt{peak}, \texttt{cpts}] = \texttt{findpeak\_opt}(\texttt{data}, \texttt{idx}, \texttt{r})$$

where `cpts` is a vector storing a 1 for each point that is a distance of $r/4$ from the path, 0 otherwise.

Image Segmentation

(a) In this section you will build upon your implementations of k-means and mean-shift to perform image segmentation. To do so, implement the function

$$\texttt{function segmIm} = \texttt{imSegment}(\texttt{im}, \texttt{p}, \texttt{alg})$$

where `im` is a color input image, p is the parameter associated with the k-means and mean-shift algorithms (i.e. p represents either k or r depending on which algorithm is run) and `alg`={'kmeans','meanshift'} specifies which clustering algorithm to use. In summary, this function is constructed by reshaping the image into RGB vectors and then clusters the resulting color data using either k-means or mean-shift depending on the value of `alg`. The segmented image is then constructed using the cluster labels and mean or peak values.

Note k-means and mean-shift both cluster using Euclidean distance metrics. As we saw earlier in the class when we studied the MacAdam ellipses of the CIE xy chromaticity diagram, Euclidean distance in RGB space does not correlate well to the perceived change in color. For example, in the green portion of the spectrum large distances are perceived as the same color, whereas in the blue part of the spectrum a small distance may represent a large change in perceived color (see Figures 6.13 and 6.14 of Forsyth and Ponce). For this reason we will use the non-linear Luv color space. In this space Euclidean distance better models the perceived change in color. In `imSegment` cluster the image data in the Luv color space by first converting the RGB color vectors to Luv using the provided MATLAB function `rgb2luv`. Then convert the resulting cluster centers back to RGB using the function `luv2rgb`.

(b) Segment the images `sunset.bmp` and `terrain.bmp` using both k-means and mean-shift with $k = 3, 5, 7, 10$ and $r = 5, 10$. For each segmentation method, display the results of each segmentation in a subplot, labelling each image with its corresponding k or r value. What effect does varying k and r have on the resulting segmentations? Compare the regions formed by k-means for different values of k. How are these regions different than those formed by mean-shift for the different values

of r? Explain. Note, although more efficient implementations exist, the above simplified implementation of mean-shift may take several minutes to run for the provided input images.