TDA 231 Machine Learning: Homework 1

Goal: Maximum Likelihood Estimation, Bayes Error rate, Classification

Due Date: January 27, 2012

General guidelines:

1. All datasets can be downloaded from the course website.
2. All matlab files have to be submitted as a single zip file named *code.zip*.
3. All plots, tables and additional information should be in a single pdf file named *report.pdf*.
4. The report should clearly indicate your group number on the fire system, your names, personal numbers and email addresses.

Useful matlab functions:

- *General:* arrayfun, cellfun, crossvalind, reshape, (anonymous functions using @), mat2cell, cell2mat
- *Plotting:* plot, scatter, legend, hold, imshow, subplot, grid, title, saveas

1. (2 points) Consider a dataset consisting of i.i.d. observations generated from a distribution $N(\mu, \sigma^2 I)$, where $\mu \in \mathbb{R}^n$, I is the $n \times n$ identity matrix, and σ^2 is a scalar.
   (a) Implement a matlab function that estimates the mean μ and variance σ^2 from the given data, using the skeleton code provided below.

```
function [mu, sigma] = sge(x)
%
% SGE Mean and variance estimator for spherical Gaussian distribution
%
% x               : Data matrix of size n x d where each row represents a
%                   d-dimensional data point e.g.
%                       x = [2 1;
%                            3 7;
%                            4 5 ] is a dataset having 3 samples each
%                            having two co-ordinates.
%
% mu              : Estimated mean of the dataset [mu_1 mu_2 ... mu_d]
% sigma           : Estimated standard deviation of the dataset (number)
%
YOUR CODE GOES HERE
```

   (b) Implement a function which takes as input a two-dimensional dataset x (as described above); and draws on the same plot the following:
      1. Scatter plot of the original data x,
      2. Circles with center μ and radius $r = k\sigma$ for $k = 1, 2, 3$ where μ and σ^2 denote the estimated mean and variance using *sge()*.
      3. Legend for each circle indicating the fraction of points (in the original dataset) that lie outside the circle boundary.

```
function [f1, f2, f3]=sol1_1(x)
%
% x: two-dimensional dataset as in sge()
%
% f1: fraction of points lying outside circle of radius 1*sigmaEst around muEst
% f2: fraction of points lying outside circle of radius 2*sigmaEst around muEst
% f3: fraction of points lying outside circle of radius 3*sigmaEst around muEst
%
YOUR CODE GOES HERE
```

(c) Run your code on the dataset dataset1.mat. Submit the resulting plot, as well as your implementation of *sge()* and *sol1_ 1()*.

2. (3 points) Dowload the dataset dataset2.mat contains 3 dimensional data generated from 2 classes with labels either $+1$ or -1. Each row of x and y contain an observation data and label respectively. There are 1000 instances of each class.

(a) Assuming that the class conditional density is spherical gaussian, use *sge*() to estimate the densities. Compute the Bayes classifier, *sph_bayes*().

```
function [Ytest]=sph_bayes(Xtest, ...) % other parameters needed.
```

(b) Let μ_1, σ_1^2 and μ_2, σ_2^2 be the means and variances of class 1 and class 2 respectively. Write a function

```
function [Ytest] = new_classifier(Xtest, mu1, mu2)
```

which implements the following classifier, we call $new\_classifier$,

$$f(x) = \text{sign}\left(\frac{(\mu_1 - \mu_2)}{\|\mu_1 - \mu_2\|}^\top (x - b)\right)$$

The parameter $b = \frac{1}{2}(\mu_1 + \mu_2)$.

(c) Report in a table 10-fold cross validation error for both classifiers with the minimum and maximum error on *dataset2*. Submit both your implementations.

3. (5 points) Download dataset digits.mat. This contains a $256 \times 1100 \times 10$ matrix *data* contains the 1100 images of handwritten digits $(1-9, 0)$ each of size 16×16. For example $data(:, 45, 3)$ is the feature vector (data) corresponding to 45^{th} sample for class 3 (corresponding to digit 3). You can visualize this using the commands

```
y = reshape(data(:, 45, 3) , 16, 16); % 16x16 image
imshow(y);                            % show image
```

(a) Use $new\_classifier$ and sph_bayes designed previously to do the classification between class 5 and 8. Report 10-fold cross validation error as well as minimum and maximum errors.

(b) Submit two images (generated using *imshow* and *subplot*) showing the mis-classified digits for each class (5 and 8) for the *sph_ bayes* classifier. Indicate on each subplot, the index of the mis-classified digit using *title* command.

(c) Investigate alternative feature functions, after scaling each pixel value between 0 and 1 instead of original gray-scale $(0 - 255)$.

1. Variance $E\left[\left(\frac{x-\mu}{\sigma}\right)^2\right]$ along each row and column of the image (32 features). For example, the variance of the scaled first row of image can be obtained as:

```
z = y(1, :)/255;        % scaled row
var_scaled = var(z);    % variance
```

2. (Optional) Skewness $E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$ and Kurtosis $E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]$ along each image row and column.

3. Divide the image into blocks of $2, 4, 8$ equal sized squares (super-pixels) by averaging the pixels values within a block. Use these as the new image.

(d) Report 10-fold cross validation results for each of the features in a single table.

The design of features is crucial part of computer vision research. Several image descriptors (features) are known and are well suited to different applications e.g. SIFT (face detection), HOG (segmentation), color histogram, etc. Often, best results are obtained by combining a number of features.