Pattern Recognition: Assignment 1

Due on Wednesday, April 25 2012, 13:00

Prof. Fred Hampprecht, Summer Term 2012

patternrecognition@hci.iwr.uni-heidelberg.de

`http://hci.iwr.uni-heidelberg.de/MIP/Teaching/pr/`

Matlab, k-Nearest Neighbor Classifier

In this exercise, you will get in touch with Matlab and implement your first classifier, k-Nearest Neighbors (kNN). You will apply it and evaluate it on the famous MNIST database of handwritten digits.

Files

```
calc_dist_mat_loop_a_b.m*
calc_dist_mat_squ_a_b.m*
displayData.m
ex01.m*
matlab_intro.m
mnist-digits.mat
```

Use ex01.m as a skeleton for your solution and fill in the missing parts in the files marked with *.

Data Description

The MNIST database of handwritten digits comprises 70,000 images. We will work on a subset of 7,000 images, split into 6,000 training samples and 1,000 test samples. You can find more information about the data at `http://yann.lecun.com/exdb/mnist/`. Our data is already converted to Matlab format and stored in the `mnist-digits.mat` file.

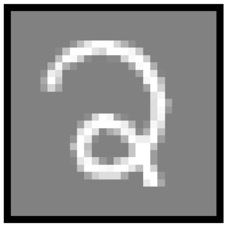
Prob. 1: Matlab

(a) Introduction to Matlab

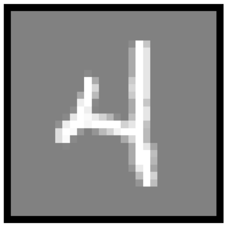
If you are new to matlab, please familiarize yourself with the most important commands (see matlab_intro.m). Feel free to skip this step, if you are a proficient matlab user.

(b) Exploring the Data (3 points)

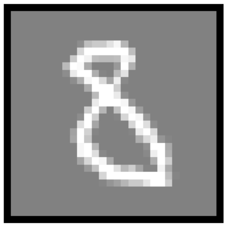
Familiarize yourself with the data description (see above). Use the matlab "load" command to load the four matrices "training", "training_label", "test", and "test_label" from "mnist-digits.mat". "training" and "test" contain one digit per row. The value of the digit (1,2,...10) can be looked up in the corresponding row in "*_label".
You can plot one or more digits using the displayData function (provided with the assignment). For example, displayData(training(201:210,:)) will plot 10 digits from the training dataset.



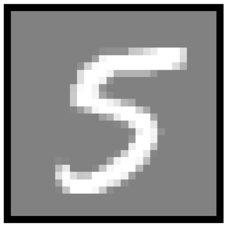
(a)



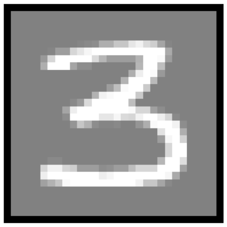
(b)



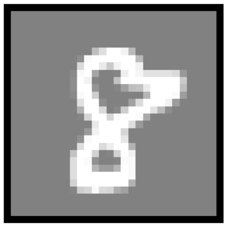
(c)



(d)



(e)



(f)

Of the above digits, which one is digit #2932 in the training dataset? Which one is digit #814 in the test dataset? How many digits in the training dataset have the label "5"?

Prob. 2: k-Nearest Neighbor Classifier

Distinguish the digit 5 from all other digits.

(a) Distance function computation using loops (3 points)

As a first step, write a matlab function dist_loop=calc_dist_mat_loop_a_b(training, test) which computes the Euclidian distance between all digits in the training and test set using loops and measure the run time (suggested commands: tic toc). The input should be the n1 x p and n2 x p training and test matrices with

Prob. 2: k-Nearest Neighbor Classifier continued on next page. . .

p pixels per image and n1 respectively n2 samples in the training resp. test set. The output should be a n1 x n2 distance matrix.

(b) Distance function computation using vectorization (4 points)

Since loops can turn out to be rather slow in matlab, and since we may be in need of efficient code later on, write a second matlab function for computing the distance function which relies on vectorization and does not have a for loop, dist_squ=calc_dist_mat_squ_a_b(training, test) and compare the run time.

(c) Write a k-nearest neighbor classifier (6 points)

Using the distance matrix, now implement a k-nearest neighbor classifier and use it to predict if a digit is the digit “5” or not for k=5. Compute the correct classification rate as the number of digits which were predicted correctly as five or notfive. (suggested commands: sort)

(d) Parameter adjustment (4 points)

Try varying the value for k (up to k=20) and compute the correct classification rate. Describe the dependency of the classification performance on k. Explain your observation.

Prob. 3 (Bonus): Error likelihood of k-nearest neighbor classification (6 points)

Compute the likelihood that a k-nearest neighbor classifier comes to a wrong conclusion due to the small size of k. Assume that for a given position in feature space, the likelihood of the (not-)five label is p(five—x)=0.4. Compute and plot the distribution of the outcome of the classification votes (0,,k for votes for “five”) for k=5 and k=20 (suggested command: factorial). How often do these classification votes lead to an erroneous conclusion? Why do you think that setting k=20 might not always be superior to setting k=5 in practice? (Note, that we assume uniform distributed labels in feature space. In real world applications labels form clusters in feature space.)

Regulations

Please hand in the matlab code, figures and explanations (describing clearly which belongs to which). Non-trivial sections of your code should be explained with short comments, and variables should have self-explanatory names. Plots should have informative axis labels, legends and captions. Please enclose everything into a single PDF document (e.g. use the publish command of MATLAB for creating a LaTeX document and run latex, dvips and ps2pdf or copy and paste everything into an office document and convert to PDF). Please email the PDF to patternrecognition@hci.iwr.uni-heidelberg.de before the deadline specified below. You may hand in the exercises in teams of two people, which must be clearly named on the solution sheet (one email is sufficient). Discussions between different teams about the exercises are encouraged, but the code must not be copied verbatim (the same holds for any implementations which may be available on the WWW). Please respect particularly this rule, otherwise we cannot give you a passing grade. Solutions are due by email at the beginning of the next exercise (April 25, 13:00).