Boston University
Department of Electrical and Computer Engineering

ENG EC 500 B1 (Ishwar) Introduction to Learning from Data

Matlab Exercise 1 Solution

Issued: Tue 22 Sep 2015 **Due:** 5pm Fri 9 Oct 2015 in box outside PHO440 + Blackboard
Required reading: Your notes from lectures and additional notes on website on classification.

- **Advise**: This homework assignment requires a large amount of time and effort. We urge you to start right away. This assignment cannot be finished by staying up all night just before the deadline.
- This homework assignment requires programming in MATLAB. If you are new to MATLAB programming, please refer to the following link for a primer:
`http://www.math.ucsd.edu/~bdriver/21d-s99/matlab-primer.html`
- You will be making two submissions: (1) A paper submission in the box outside PHO440. (2) An electronic submission of all your MATLAB code (in a single zipped file appropriately named as described below) to blackboard.
- **Paper submission:** This must include all plots, figures, tables, numerical values, derivations, explanations (analysis of results and comments), and also printouts of all the MATLAB .m files that you either created anew or modified. Submit color printouts of figures and plots whenever appropriate. Color printers are available in PHO307 and PHO305. Be sure to annotate figures, plots, and tables appropriately: give them suitable *titles* to describe the content, label the *axes*, indicate *units* for each axis, and use a *legend* to indicate multiple curves in the plots. Please also explain each figure properly in your solution.
- **Blackboard submission:** All the MATLAB .m files (and only .m files) that you either create anew or modify must be appropriately named and placed into a **single** directory which should be zipped and uploaded into the course website. Your directory must be named as follows: <yourBUemailID>_matlab1. For example, if your BU email address is `charles500@bu.edu` you would submit a single directory named: `charles500_matlab1.zip` which contains all the MATLAB code (and only the code).
- **File naming convention:** Instructions for file names to use are provided for each problem. As a general rule, each file name must begin with your BU email ID, e.g., `charles500_<filename>.m`. The file name will typically contain the problem number and subpart, e.g., for problem 1.1b, the file name would be `charles500_matlab1_1b.m`. Note that the dot . in 1.1 is replaced with an underscore (this is important).

Problem 1.1 Gaussian Discriminant Analysis:
In this problem we consider using various Gaussian Discriminant Analysis classifiers.

(a) Write 4 generalized MATLAB "helper" functions for training and testing a Quadratic Discriminant Analysis (QDA) classifier and a Linear Discriminant Analysis (LDA) classifier. We have provided 4 template MATLAB function files (which specifies what inputs and outputs you should use) for you to write, complete, and **submit**:

1. `QDA_train.m`: outputs a trained QDA classifier given input training data and labels
2. `QDA_test.m`: outputs labels given a trained QDA classifier and test data
3. `LDA_train.m`: outputs a trained LDA classifier given input training data and labels
4. `LDA_test.m`: outputs labels given a trained LDA classifier and test data

Rename these helper functions to: <yourBUemailID>_<helper file name>.m. Example, `charles500_QDA_train.m`. Comment your code as necessary. Do **not** use built-in MATLAB functions for discriminant analysis such as `fitcdiscr`.

Dimension of samples $\gg$ *the dimension of features*: Load the dataset provided in `data_iris.mat`. In this dataset, the rows of variable `X` correspond to samples (150 total), and the columns correspond to feature dimension (4 total). The elements of variable `Y` are the ground truth class labels for each data sample.

(b) Classification involves the following steps (1) splitting the data into a training set (in this problem 100 samples) and a test set (the remaining 50 samples), (2) learning a classifier on the training set, (3) predicting the class labels of the samples in the test set, and (4) calculating a performance metric.

Create one training set by choosing 100 samples (picked uniformly at random without replacement) from the entire dataset. Repeat this 10 times for 10 different training/test splits. For each split, conduct the remaining 3 steps (training, testing, performance evaluation). Calculate the Correct Classification Rates (CCR) and the Confusion Matrices using the generalized QDA and LDA functions you have implemented in part (a) for each training/test split.
Report the following results:

- Mean vectors of training samples for each class (common to both LDA and QDA), averaged over 10 splits.
- The variances of all the 4 dimensions (the diagonal terms of the covariance matrix) for each class of training set (in QDA), averaged over 10 splits.
- The overall variances of all the 4 dimensions (in LDA), averaged over 10 splits. **Note:** all classes share the same covariance matrix in LDA.
- The mean and standard deviation of all 10 test CCRs.
- The confusion matrices of the two splits which have the best and the worst CCR (in LDA).
- **Submit** your script with name <yourBUemailID>`_matlab1_1b.m` that generates all above results.

Helpful MATLAB functions: `randperm, confusionmat`

Dimension of data samples $\ll$ *dimension of feature*: Load the dataset provided in `data_cancer.mat`. Rows of variable `X` correspond to data samples (216). The elements of variable `Y` are the corresponding ground truth class labels. In this dataset, the dimension of features (4000) is much larger than the number of samples (216). As a result, the estimates of the covariance matrices $\widehat{\Sigma}$ in LDA/QDA will be singular. One strategy to account for this is to consider a *Regularized* Discriminant Analysis (RDA) in which the common covariance matrix in LDA is replaced by,

$$\widehat{\Sigma}_{\text{reg}} = \lambda \text{diag}(\widehat{\Sigma}_{\text{LDA}}) + (1 - \lambda)\widehat{\Sigma}_{\text{LDA}} \tag{1}$$

where $\lambda \in [0, 1]$ controls the amount of regularization and $\text{diag}(\mathbf{A})$ is a matrix whose diagonal entries are the same as $\mathbf{A}$ where the off-diagonal entries are 0.

(c) Implement your code for RDA using the provided "helper" function templates `RDA_train.m` and `RDA_test.m`. Comment your code and submit the completed MATLAB functions. As before, rename these helper functions to: `<yourBUemailID>_<helper file name>.m`.
Helpful MATLAB function: `diag`

(d) Split the data into a training set (of size 150) and use the remaining as test data (of size $216-150 = 66$). Choose the 150 training samples uniformly at random from the entire dataset as you did in part (b). For each choice of λ in the range of $[0.1 : 0.05 : 1]$ (x-axis), calculate the training and test CCR (y-axis) and report them on the same plot. **Submit** your script with name `<yourBUemailID>_matlab1_1d.m` along with its corresponding figure. Make sure that if the grader runs your code, it will generate the exact same results that you are reporting. This can ensured by fixing the random seed. See MATLAB function: randstream.

Solution:

(a) See the example code provided.

(b) See the example code provided.

- The mean vectors are
  $\mu_{1,\text{averaged}} = [5.0162, 3.4473, 1.4649, 0.2491]$ for class 1
  $\mu_{2,\text{averaged}} = [5.9300, 2.7427, 4.2393, 1.3109]$ for class 2, and
  $\mu_{3,\text{averaged}} = [6.5939, 2.9662, 5.5431, 2.0205]$ for class 3.

- Averaged attribute variances for 3 classes in QDA are
  $0.1181, 0.1321, 0.0326, 0.0115$ for class 1,
  $0.2608, 0.0973, 0.2212, 0.0383$ for class 2,
  $0.3709, 0.0996, 0.3069, 0.0746$ for class 3.

- The overall attribute variances in LDA are $0.2498, 0.1099, 0.1865, 0.0413$.

- Average test CCR is 97.6% for QDA and 97.8% for LDA. The standard deviation of CCRs is 1.8% for QDA and 2.0% for LDA.

- For LDA, the confusion matrix with the best test CCR is

	1	2	3
1	15	0	0
2	0	20	0
3	0	0	15

  and the confusion matrix with the worst test CCR is

	1	2	3
1	17	0	0
2	0	16	1
3	0	2	14

(c) See the example code provided.

(d) See the example code provided. The training and test CCR plots should be similar to the ones shown in Figure 1. Notice that the training CCR is close to one for all nonzero λ's (at least for λ's larger

than 0.1) and drops significantly when λ is very close to one. This implies that that the classes can be well-separated by the regularized LDA model. When $\lambda = 1$ (extreme regularization), the regularized covariance estimate becomes diagonal and the model looses all ability to capture dependencies between individual attributes resulting in a significant drop in CCR. If $\lambda = 0$ (no regularization) on the other hand, the covariance estimate will become singular and the method will break down (not shown in the plot). The test CCR follows a similar trend. The best test CCR value is attained for a range of λ's close to one. The exact values of the best test CCR and the associated best λ will depend on the training-test split which was done randomly at the beginning. If one were to repeat this experiment for different random training-test splits and average the resulting CCR curves, one would find a similar trend: poor CCR at the extremes (for different reasons though) with the best performance in the middle. This example also illustrates that just a little bit of regularization may be enough to stabilize the performance (at least for this dataset).

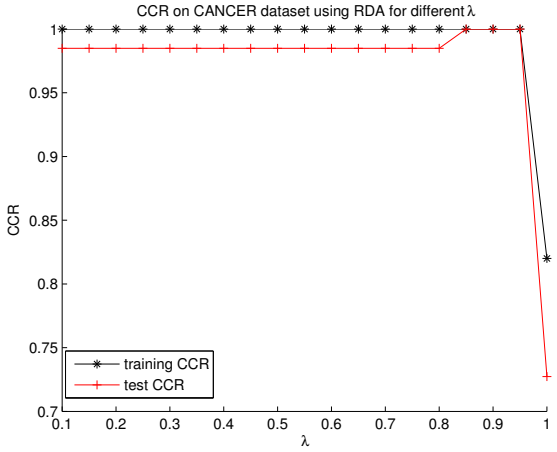

Figure 1: *CCR plots for Problem 1(d).*

Problem 1.2 Naive Bayes Text Document Classifiers: In this problem we classify text documents using Naive Bayes. We use the classic *20 newsgroup* dataset. Here, each document is labeled as one of 20 different classes and is provided in `data_20news.zip`. This directory contains the following files:

1. `vocabulary.txt`: a list of all the words that can possibly appear in the documents. The line number of a word is its ID (Word_ID).

2. `newsgrouplabels.txt`: the names of the 20 classes. The line number of a class (label) is its ID.

3. `train.data` and `test.data`: the words in each of the documents. Each line of each file is of the form: "Document_ID, Word_ID, Word_count". Word_count is the number of times word Word_ID appears in document Document_ID.

4. `train.label` and `test.label`: the labels of the training and test documents.

5. `stoplist.txt`: a list of so-called "stop words" such as "the", "is", "on", etc. This will be used only in part (f).

Let $\{1, \ldots, W\}$ denote the W distinct words in the vocabulary. We view each document $\mathbf{x}_j$, $j = 1, \ldots, n$ as a collection of words $\{w_{1,j}, \ldots, w_{d_j,j}\}$ where d_j is the number of words in document $\mathbf{x}_j$ and y_j is the label of

the document. Each word $w_{i,j}$ takes values in $\{1, \ldots, W\}$. The Naive Bayes classifier assumes that:

$$P(Y_j = c|\mathbf{x}_j) \propto P(Y_j = c) \prod_{i=1}^{d_j} P(w_{i,j}|Y = c), \quad c = 1, \ldots, 20 \tag{2}$$

We denote $P(Y_j = c) = \pi_c, c = 1, \ldots, 20$ and $P(w_{i,j} = w|Y = c) = \beta_{w,c}, w = 1, \ldots, W, c = 1, \ldots, 20$. Note that by definition, $\sum_{w=1}^{W} \beta_{w,c} = 1$, for each $c = 1 \ \ldots, 20$.

(a) Write a MATLAB script to load and parse the data. Name this file <yourBUemailID>_matlab1_2a.m. **Report** the following statistics

- The total number of **unique** (i.e., distinct) words that appear in the training set, the test set, and the entire dataset, respectively.
- The average document length (in terms of number of words) in the training and test sets, respectively.
- The total number of unique words that appear in the test set, but not in the training set.

Helpful MATLAB functions: `textread, sparse`.

We next train a Naive Bayes classifier. We first learn the parameters $\beta_{w,c}$'s using Maximum Likelihood Estimation (MLE):

$$\beta_{w,c} \propto n_{w,c} \tag{3}$$

where $n_{w,c}$ is the total number of the word w that appears in documents with the label c.

(b) Train a Naive Bayes classifier using the MLE rule from Eq. 3 and the training set. Afterwards, evaluate this classifier on the test set. Do not use the built-in MATLAB function for the Naive Bayes classifier. **Submit** your script with name <yourBUemailID>_matlab1_2b.m. **Report** the following results:

- Among the $W \times 20$ estimated parameters (the $\beta_{w,c}$'s), how many of them are non-zero?
- For some test documents, $P(Y = c|\mathbf{x}) = 0$ for all $c = 1, \ldots, 20$. What is the total number of such test documents? Can you explain why their probabilities are zero?
- The test CCR.

(c) One strategy to overcome the issue encountered in part (b) is to remove words that only appear in the test documents (but not in the training documents). Remove these words. Repeat training and testing using the MLE rule in (b) and **report:**

- The total number of non-zero estimated $\beta_{w,c}$'s.
- The test CCR.

To learn the parameters $\beta_{w,c}$'s we can also use the Maximum Aposteriori Probability (MAP) rule. It is common practice to impose a Dirichlet prior on the $\beta_{w,c}$'s and as such, Eq. 3 becomes:

$$\beta_{w,c} \propto n_{w,c} + \alpha \tag{4}$$

for some small $\alpha > 0$.

(d) Train a Naive Bayes classifier using the MAP rule in Eq. 4 with $\alpha = 1/W$, and evaluate it on the test set. Do not use the built-in MATLAB function for the Naive Bayes classifier. **Submit** your script with name <yourBUemailID>_matlab1_2d.m. **Report** the following results:

- The test CCR.
- The 20×20 confusion matrix. Make sure that you annotate the confusion matrix appropriately.

(e) Calculate the test error for different choices of α's in the range $10^{-5}, 10^{-4.5}, 10^{-4}, 10^{-3.5}, ..., 10^{1}, 10^{1.5}$. Plot the test CCRs (y-axis) as functions of α (x-axis). Use a log-scale for α.

(f) In classification tasks with text observation, the most common words (*stop words*) such as "the","is", etc., may be less informative. A common practice is to remove such stop words. The file `stopword.txt` contains a list of such stop words. Remove these words from all the document and the vocabulary and then repeat the steps in part (d). **Report:**

- The size of the new dictionary (the new value of W) after removing the stop words.
- The average document length (in terms of number of words) in the training and test sets respectively.
- The test CCR.

Solution:

(a) See the provided example code for parsing the data.
• Total number of unique words in training set: 53975, in test set: 47376, in the entire dataset: 61188.
• Average document length in training set: 245.39, in test set: 239.43.
• 7213 unique words appeared only in test set.

(b) See the provided example code.
• The number of non-zero parameters $\beta_{w,c}$ is 200778 which is 16.41% of all the parameters.
• 6958 test documents are predicted to have zero probability for all 20 classes. These are the test documents which contain at least one word which does not appear in any training document.
• The test CCR should be around 9.46%.

(c) See the provided example code.
• The number of non-zero parameters $\beta_{w,c}$ is 200778 which is 18.6% of all the parameters.
• The test CCR should be around 11.14%.
These low CCR values show that the scarcity of data relative to the number of parameters can have a major impact on the classification performance. This motivates the use of regularization in learning.

(d) See the provided example code. You will find a huge improvement in the classification results.
• Test CCR is 78.52%.
• The confusion matrix is in Table 1

(e) See Figure 2. The trend is similar to Problem 1(d): poor CCR at the extremes with the best performance in the middle. The best CCR is attained of α slightly less than one (recall that the horizontal axis is on a logarithmic scale). The CCR seems to be quite stable for a wide range (on a logarithmic scale) of very small values of α, but seems to drop dramatically for $\alpha > 1$. This is because most words have relatively small counts in any document. Since α acts as a default count, if $\alpha \gg 1$, it starts becoming the dominant count rendering all class conditional pmfs close to uniform and the classification performance close to random guessing.

(f) See the provided example code.
• The size of the pruned vocabulary is 60698.
• The average document length in the training set: 116.98, in the test set: 114.62

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	249	0	1	0	0	0	0	0	0	0	2	0	3	5	0	11	1	12	6	39
2	0	286	33	11	17	54	7	3	1	0	0	3	20	7	8	2	1	1	1	3
3	0	13	204	30	13	16	5	1	0	0	1	0	4	0	0	0	0	0	0	0
4	0	14	57	277	30	6	32	2	1	1	0	3	25	3	1	0	0	1	0	0
5	0	9	19	20	269	3	16	0	0	1	0	4	7	0	0	0	0	0	1	0
6	1	22	21	1	0	285	1	0	0	0	0	1	4	0	3	2	1	0	1	0
7	0	4	4	10	12	1	270	14	2	2	2	0	8	3	1	1	1	1	0	0
8	0	1	2	2	2	1	17	331	27	1	1	0	11	5	0	0	2	2	0	0
9	1	1	3	1	2	3	8	17	360	2	2	0	6	4	1	0	1	0	0	1
10	0	0	0	0	0	0	1	0	0	352	4	1	0	1	0	0	1	2	0	1
11	0	1	0	1	0	0	2	0	0	17	383	1	0	0	1	0	0	0	0	0
12	2	11	12	4	3	5	0	1	0	0	0	362	21	1	4	0	4	2	5	1
13	0	8	5	32	21	3	7	13	3	1	0	2	264	8	6	0	0	1	0	0
14	3	6	10	1	8	6	4	0	1	3	0	2	9	320	5	2	5	0	10	2
15	3	10	8	2	4	4	6	4	0	3	0	2	7	8	343	0	2	0	6	6
16	24	1	3	0	0	0	0	2	0	5	1	0	1	7	3	362	1	6	2	27
17	2	2	1	0	1	1	2	0	1	2	2	9	3	6	2	0	303	3	63	10
18	3	0	0	0	0	1	1	0	1	1	0	0	0	5	1	1	5	326	6	3
19	4	0	5	0	1	1	2	6	0	5	1	5	0	8	12	2	23	18	196	7
20	26	0	3	0	0	0	1	1	0	1	0	0	0	2	1	15	13	1	13	151

Table 1: *Confusion matrix of the Bayesian Naive Bayes classifier on the 20 Newsgroup dataset (Problem 2(d)). Columns represent the ground truth and the rows the decisions.*

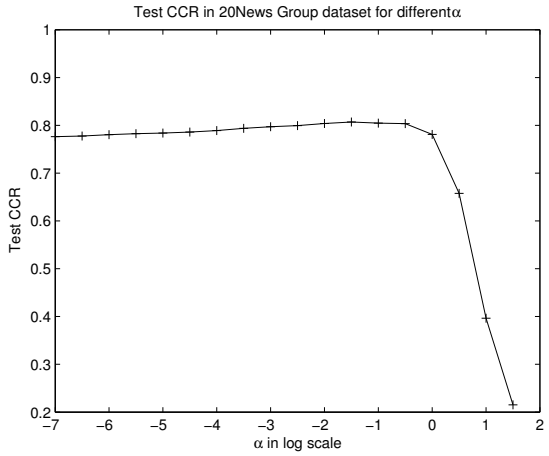

Figure 2: *Solution plot for 2(e).*

• The test CCR is 78.23%. The test CCR is similar to when stop words are not removed. This indicates that in Naive Bayes classification, the stop words have a limited impact on the classification performance. This is because stop words can be expected to "equally" affect all classes in a similar way.

Problem 1.3 Nearest Neighbor Classifier: In this problem we will apply a k- Nearest Neighbor classifier based on the Euclidean distance to two datasets. As in previous problems, you are not allowed to use the built-in MATLAB functions for nearest neighbor classification. First, we consider a simulated data set provided in `data_knnSimulation.mat`. The rows of variable `Xtrain` are data points and each data point has 2 feature attributes.

(a) Create a scatter plot of all the training data (with the first column as the x-axis and the second column as the y-axis). Color the data points in the 1st, 2nd, and 3rd classes with red, green, and blue colors, respectively.
Helpful MATLAB function: `gscatter`

(b) For each point on this 2-D grid $[-3.5 : 0.1 : 6] \times [-3 : 0.1 : 6.5]$, calculate its probability of being class 2 using $k = 10$ nearest neighbors. Plot the probabilities of these points on a 2-D colored map using the MATLAB default colormap. Repeat this for class 3. **Submit** the two figures. Add a colorbar in the figures to indicate the color range.
Helpful MATLAB functions: `imagesc, contourf, colormap, colorbar`.

(c) For all the points on the same 2-D grid used in part (b), predict its class label using a k-NN classifier with $k = 1$. Color-code the decisions for each grid point using the same color coding scheme used in part (a). Repeat this for $k = 5$. **Submit** the two figures. Comment on their differences.

We next look at the classical handwritten digit recognition problem. The dataset is provided in `data_mnist_train.mat` and `data_mnist_test.mat`. Each data point represents a 28×28 grayscale image for hand written digits from 0 to 9. To visualize this data, for example, the 200-th training image, you can use `imshow(reshape(X_train(200,:), 28,28)')`.

(d) Apply a 1-Nearest Neighbor classifier to this dataset. **Report** the test CCRs and the confusion matrix. **Submit** your script with name `<yourBUemailID>_matlab1_4d.m`. Your script should be able to display in the command window the CCR that you reported.
Hint: Try to make your code time efficient. The dataset is large.

Solution:

(a) See Figure 3.

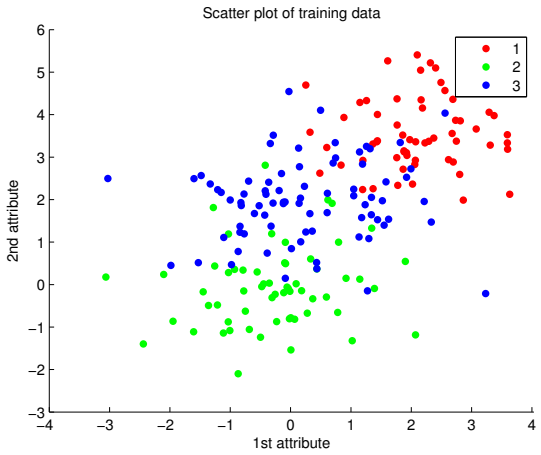

Figure 3: *Scatter plot for Problem 3(a).*

(b) See Figure 4

(c) See Figure 5. When $k = 1$ the decision boundaries are rough (noisy) with many isolated islands. When $k = 5$, the decision boundaries are smoother and more contiguous. A larger value of k has the effect of regularizing (smoothing-out) the decision boundaries.

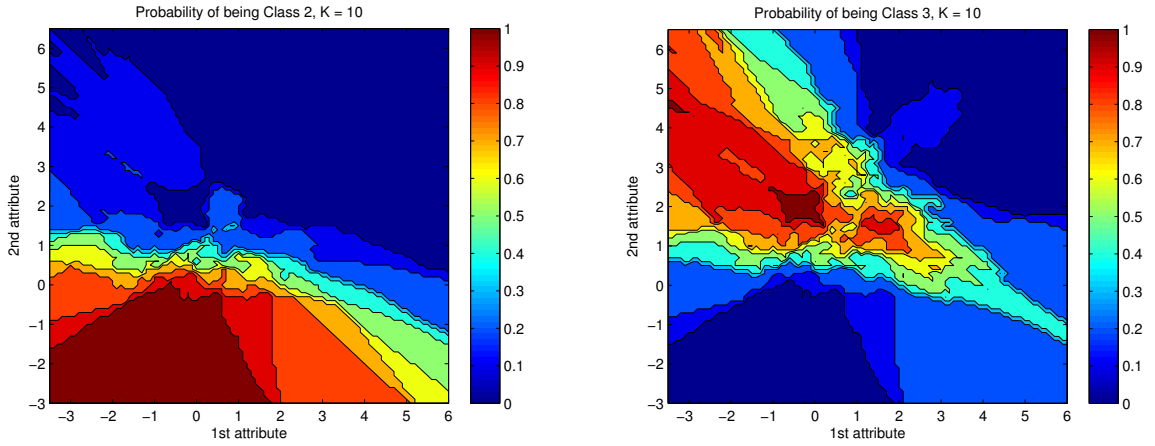

Figure 4: *Posterior class probability estimate maps using $k = 10$ nearest neighbors (Problem 3(b)).*

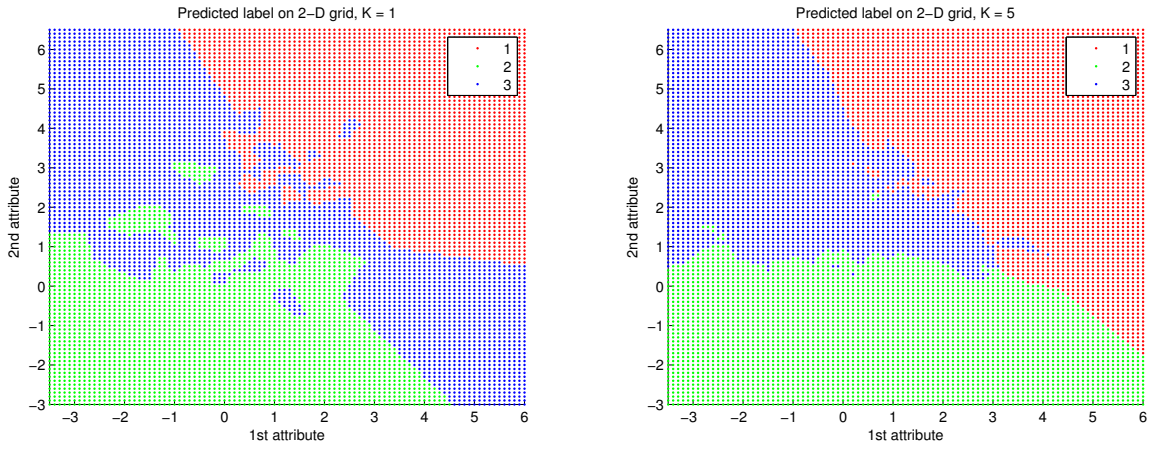

Figure 5: *Predicted class label maps using $k = 10$ nearest neighbors (Problem 3(c)).*

(d) See the example code. Euclidean distances can be computed via inner products: $\|\mathbf{x} - \mathbf{x}'\|_2^2 = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}', \mathbf{x}' \rangle - 2\langle \mathbf{x}, \mathbf{x}' \rangle$. Thus, is $\mathrm{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathrm{diag}(A)$ = column vector of the main diagonal elements of square matrix A, and $\mathbf{1}$ = column vector of all ones, then the pairwise Euclidean distance matrix (EDM) is given by:

$$\mathrm{EDM}(\mathrm{X}) = \mathbf{1}\mathrm{diag}(\mathrm{X}^\top \mathrm{X})^\top - 2\mathrm{X}^\top \mathrm{X} + \mathrm{diag}(\mathrm{X}^\top \mathrm{X})\mathbf{1}^\top$$

$$\mathrm{EDM}(i, j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2.$$

The overall CCR is 96.91%. The confusion matrix appears in Table 2

0	0	1	2	3	4	5	6	7	8	9
1	973	0	7	0	0	1	4	0	6	2
2	1	1129	6	1	7	1	2	14	1	5
3	1	3	992	2	0	0	0	6	3	1
4	0	0	5	970	0	12	0	2	14	6
5	0	1	1	1	944	2	3	4	5	10
6	1	1	0	19	0	860	5	0	13	5
7	3	1	2	0	3	5	944	0	3	1
8	1	0	16	7	5	1	0	992	4	11
9	0	0	3	7	1	6	0	0	920	1
10	0	0	0	3	22	4	0	10	5	967

Table 2: *Confusion table for the 1-nearest neighbor classifier on the MNIST dataset (Problem 3(d)). Columns represent the ground truth and the rows the decisions.*