Project 1 extra description:

1. After you loaded the dataset, you should have 4 matrices, trX, trY, tsX, tsY. For trX and tsX, each row represents a digit, you can see that their dimension is 28 * 28 = 784
2. The number of rows represents how many digits there are, so trX should have 6265+5851 rows, tsX should have 1028 + 974 rows.
3. trY and tsY are labels of training set and testing set respectively.
4. For feature extraction, you have to calculate the mean and s.d., of each digit. After the extraction, you should have a new trX and new tsX. Let’s call them trX_new and tsX_new, BOTH of them have only two columns, instead of 784.
5. You should use trX_new and tsX_new for BOTH naive bayes and logistic regression.

For Naive Bayes:

1. First separate trX_new into two sets, one set contains only digit 7, another contains only digit 8.
2. Calculate the mean and covariance matrix, for each of these sets.
3. The mean of the features are in 2d, that is, [mean of (mean of the pixel values), mean of (s.d. Of the pixel values)]. The (mean of the pixel values) and (s.d. Of the pixel values) are the features of every digit.
4. The covariance matrix should be a diagonal matrix, as you are doing naive bayes, the two features should be independent, which means the upper right and lower left element of the matrix should be 0, i.e. cov(X1, X2) = cov(X2, X1) = 0. Here, X1 is the first feature (mean of the pixel values) and X2 is the second feature (s.d. Of the pixel values).
5. Use them to calculate the required probability.

For logistic regression:

1. Just calculate the decision boundary by using gradient.

For the submission, you should upload a report in .pdf and at least one .py file (or .m if you are using matlab)
For the report, it contains several parts. The following is an example from a student last semester:

Density Estimation and Classification

[redacted]

[redacted]asu.edu

1.INTRODUCTION

MNIST dataset contains 70,000 images of [redacted]

2.FEATURE EXTRACTION

The extracted dataset contained training and testing samples for the digit "7" and "8".
Training Samples: "7": 6265; "8":5851
Test Samples: "7": 1028; "8":974

[redacted]

3.Naive Bayes Classification

To apply the Naive Bayes Classification, the two features: mean and standard deviation were [redacted]

As the probability for class 7 is higher so the sample **a** belongs to class 7.

4.Logistic Regression

The logistic regression is used for the classification task and it uses the form of sigmoid to [redacted]

3. Find the difference between the predicted value and the real value of y

[redacted]

Gradient ascent [redacted]

5. Results

The following are the results of the parameter estimation for samples 7 and 8. They contain the [redacted]

Accuracy	**Total Samples**	**Samples of 7**	**Samples of 8**
[redacted]	[redacted]	[redacted]	[redacted]