Ira A. Fulton Schools of Engineering
Arizona State University

CSE 575: Statistical Machine Learning

Project 1: Density Estimation and Classification

Purpose

In this project, we will systematically implement and examine the three major categories of Machine Learning techniques of this course, including supervised learning, unsupervised learning, and deep learning.

Technology Requirements

The specific algorithmic tasks you need to perform for this part of the project include:

1. Extracting the features and then estimating the parameters for the 2-D normal distribution for each digit, using the training data. Note: You will have two distributions, one for each digit.
2. Use the estimated distributions for doing Naïve Bayes classification on the testing data. Report the classification accuracy for both "0" and "1" in the testing set.

Algorithms:

- MLE Density Estimation, Naïve Bayes classification

Resources:

- A subset of MNIST dataset, download either from http://yann.lecun.com/exdb/mnist/ (requiring you to extract data corresponding to digit 0 and digit 1 only), or from the .mat files provided.

Workspace:

- Any Python programming environment

Software:

- Python environment

Language(s):

- Python

Directions

Lab: Project 1 Naive Bayes Classifier

Preparation

First, when you open your lab, you will be in the 'renameNB' Jupyter notebook. As you run the code, you will load the trainset and testset for digit0 and digit1 respectively (Please read code and you will understand). Both trainset and testset are sub-dataset from the MNIST dataset. The MNIST dataset contains 70,000 images of handwritten digits, divided into 60,000 training images and 10,000 testing images. We use only a part of images for digit “0” and digit “1” in this question.

Therefore, we have the following statistics for the given dataset:

- Number of samples in the training set: "0": 5000 ;"1": 5000.

- Number of samples in the testing set: "0": 980;   "1": 1135

We assume that the prior probabilities are the same (P(Y=0) = P(Y=1) =0.5), although you may have noticed that these two digits have different numbers of samples in testing sets.

In the existing code, myID is a 4-digit string and please change this string to the last 4-digit of your own studentID; train0 is your trainset for digit0; train1 is your trainset for digit1; test0 is your testset for digit0; and test1 is your testset for digit1. They are all Numpy Arrays. You can also convert them into python arrays if you like.

Other than the string named 'myID', **please do not** change any existing code and just write your own logic with the existing code.

You may go to the original MNIST dataset (available here http://yann.lecun.com/exdb/mnist/) to extract the images for digit 0 and digit 1, to form the dataset for this project. To ease your effort, we have also extracted the necessary images, and store them in “.mat” files. You may use the following piece of code to read the dataset:

- import scipy.io
- Numpyfile= scipy.io.loadmat(‘matlabfile.mat’)

Files for you to download: “**CSE 575_Project 1 Mat Files**” (attached in the project description in the course)

Programming

For your own code logic, you have 4 tasks to do:

Task 1:

You need to first extract features from your original trainset in order to convert the original data arrays to 2-Dimensional data points.

You are required to extract the following two features for each image:

- **Feature1**: The average brightness of each image (average all pixel brightness values within a whole image array)

- **Feature2**: The standard deviation of the brightness of each image (standard deviation of all pixel brightness values within a whole image array)

We assume that these two features are independent and that each image is drawn from a normal distribution.

Task 2:

You need to calculate all the parameters for the two-class naive bayes classifiers respectively, based upon the 2-D data points you generated in Task 1 (In total, you should have 8 parameters).

- (No.1) Mean of feature1 for digit0
- (No.2) Variance of feature1 for digit0
- (No.3) Mean of feature2 for digit0
- (No.4) Variance of feature2 for digit0
- (No.5) Mean of feature1 for digit1
- (No.6) Variance of feature1 for digit1
- (No.7) Mean of feature2 for digit1
- (No.8) Variance of feature2 for digit1

Task 3:

Since you get the NB classifiers' parameters from Task 2, you need to implement their calculation formula according to their Mathematical Expressions. Then you use your implemented classifiers to classify/predict all the unknown labels of newly coming data points (your test data points converted from your original testset for both digit0 and digit1). Thus, in this task, you need to work with the testset for digit0 and digit1 (2 Numpy Arrays: test0 and test1 mentioned above) and you need to predict all the labels of them.

Note: Remember to first convert your original 2 test data arrays (test0 and test1) into 2-D data points as exactly the same way you did in Task 1.

Task 4:

In Task 3 you successfully predicted the labels for all the test data, now you need to calculate the accuracy of your predictions for testset for both digit0 and digit1 respectively.

Submission Directions for Project Deliverables

Result Submission

Quiz: Project 1 Density Estimation and Classification Result Submission

As the result from your Jupyter Notebook of Project 1, you should have 8 components for computed parameters and 2 components for accuracy. The order of these 10 components should look like this:

[Mean_of_feature1_for_digit0, Variance_of_feature1_for_digit0

Mean_of_feature2_for_digit0, Variance_of_feature2_for_digit0

Mean_of_feature1_for_digit1, Variance_of_feature1_for_digit1

Mean_of_feature2_for_digit1, Variance_of_feature2_for_digit1

Accuracy_for_digit0testset, Accuracy_for_digit1testset]

When you get all the components for your answer, please go to "**Quiz: Project 1 Density Estimation and Classification Result Submission**" and do the submission there.

Note: You **cannot** submit anything in your Jupyter Notebook. In this project, Jupyter Notebook is simply a uniformed IDE we provide you to implement the project.

Report Submission

Graded Assignment: Project 1 Density Estimation and Classification Report Submission

Please write your report, including your final 8 parameters and 2 accuracies in a .pdf file. Put your code and your report together in only one .zip file and submit it. Do not forget to write down your studentID and you should have only one .zip file. Please feel free to share any of your experiences during the implementation with us if you like.

Please submit your report regarding Project 1 to the item titled "**Graded Assignment: Project 1 Density Estimation and Classification Result Submission**".

- Acceptable file types: .pdf or .doc/docx.
- Length of the report: no more than 2 A4 pages.
- Content: (The following must be included)
  - Please include the parameters estimated as well as the accuracy in the report.
  - Your observation and analysis about the project.

Evaluation

Result Submission

- 1 point for mean and variance of f1 for digit0
- 1 point for mean and variance of f2 for digit0
- 1 point for mean and variance of f1 for digit1
- 1 point for mean and variance of f2 for digit1
- 2 points for predicting new labels for digit0testset and calculating the accuracy.
- 2 points for predicting new labels for digit1testset and calculating the accuracy.

Note: The **acceptable** range for parameters is [x-0.2, x+0.2]; The **acceptable** range for accuracy is [x-0.005, x+0.005]. It means that if one of your float-number answers falls into its corresponding range, your answer will be graded as correct. No, otherwise.

Report Submission

- 1 point for the parameters estimated and the accuracy
- 1 point for the analysis