Computer Vision

SSI Pattern Recognition

Practical

Désiré Sidibé
dro-desire.sidibe@u-bourgogne.fr

Linear Classification

I. Introduction

In this lab, we will implement linear discriminant analysis and logistic regression.
We will try to recognize images.
You shall submit a clearly written and commented report as well as your own code.

II. Female or Male?

The file `heightWeightData.txt` contains some data showing the height and weight of different people. The first column is the class label (1=male, 2=female), the second column is the height and the third is the weight.

1. Extract the height/weight data corresponding to males. Let $\mathbf{X}$ be the corresponding $N \times 2$ matrix, where N is the number of men. Fit a 2d Gaussian to the male data using the empirical mean and covariance. Plot your Gaussian as en ellipse and superimpose it on your scatter plot of the data.

   Recall that empirical mean and covariance are given by

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \bar{\mathbf{x}}, \quad \text{and} \quad \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

2. Standardize the data (i.e. make each feature have 0 mean and unit variance) and re-estimate the 2d Gaussian.

   Recall standardization : $x_{ij} \leftarrow \frac{x_{ij} - \bar{x_j}}{\sigma_j}$, in each dimension j.

3. Compute the eigenvectors $\mathbf{U}$ and eigenvectors $\mathbf{\Lambda}$ of $\mathbf{X}$. Then we can transform the data by computing $\mathbf{\Lambda}^{-1}\mathbf{U}^T\mathbf{x}$ for each data vector $\mathbf{x}$. This called data *whitening* or *sphering*.

   - Apply data whitening and re-estimate the 2d Gaussian
   - What is the difference with data standardization? Explain?

We now want to use the height/weight features to classify each person as male or female.

1. Use LDA to perform this classification and compute the training error. Note that we will use all data as training set.

2. We can also use PCA to project the data onto a 1d space. Compute the corresponding principal component and explain the difference with the decision boundary found by LDA?

III. Digits recognition

You are gievn a set of handwritten digits in the form of binary images. Each digit correspond to a 28×28 image and is represented by a $28 * 28 = 784$ dimensional vector.
The file `trainingData.mat` contains a set of training digits images with their corresponding lables (from 0 to 9), while `testData.mat` contains a set of test data.
Your task is to build a classifier, using the training data, and use it to correctly predict the labels of the test data.

1. **Data visualization**:
   You can visualize some of the training data, by transforming the digits back to images! That is, each image is represented by a 784 vector which can be reshaped into a 28×28 matrix for displaying.

2. **Dimension reduction**:
   First of all, we see data each digits is a high dimensional vector, $D = 784$, and the values of this vector are highly correlated.

   (a) Use PCA to reduce the dimension of the data, keeping only the first two principal components ($d = 2$).

   (b) Visualize some training examples in the 2D eigenspace.

3. **Classification**:
   As we have multiple classes, exactly $K = 10$ in this problem, we will use a *one-vs-all* classification strategy.
   The idea is to build K bianry classifiers. For example, for class k, we use all training example that belong to class k as positive examples, and consider all other training data as negative examples. We can than learn a binary classifier f_k for class k. Then, we classify new data as

   $$f(\mathbf{x}) = \arg \max_k f_k(\mathbf{x}).$$

   (a) Try regularized logistic regression and repport the training and testing classification error.

   (b) Try SVM and repport the training and testing classification error.

4. We have reduced the dimesion to 2, but this was an arbitrary choice. How can we select a better value for d?
   Explain your answer and show the results obtained.