Project 1: Density Estimation and Classification Report

CSE 575: Statistical Machine Learning (Summer 2024) (Samira Ghayekhloo)

Group 2

Project Team: Arjun Dadhwal, Ching-Chun Yuan, Jeffrey Li

Table of Contents

Parameter estimation

The two features extracted from the samples in both the training set and the testing set were the mean and standard deviation of the brightness of all the pixels for each of the samples (images).

The two parameters were estimated for the 2-D normal distributions for both digits (0 and 1), using the training data for each.

Parameter 1: Mean

The first parameter of the 2-D normal distribution was the Mean vector μ for both the features of all images.

The mean vector had two component: , μ_1, the mean of the values of the first feature for each of the sample, i.e. the mean value of the brightness of all the sample image's pixels, and μ_2, the mean of the values of the second feature for each of the sample, i.e. the standard deviation of the brightness of all the sample image's pixel from their mean. It can be represented as:

$$\mu = [\mu_1, \mu_2]$$

The components μ_1 and μ_2 are calculated using given formula:

$$\mu_1 = \frac{\sum_{i=1}^{n} x_{1i}}{n} \text{ and } \mu_2 = \frac{\sum_{i=1}^{n} x_{2i}}{n}$$

Where μ_1 is the first component of the mean, μ_2 is the second component of the mean, x_{1i} is the value of the first feature for the ith sample, x_{2i} is the value of the second feature for the ith sample, n is the total number of samples.

Estimated value of parameter Mean

For Digit 0 training set

The estimated value of the mean = [44.21682791 87.43729857]

For Digit 1 training set

The estimated value of the mean = [19.37965385 61.36879749]

Parameter 2: Covariance Matrix

The second parameter for the 2-D normal distribution was the Covariance matrix for both the features. It is represented as:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Where σ_1^2 represents the variance of the values of the first feature of all samples, σ_{12} represents the covariance between first and the second feature, σ_2^2 variance of the values of the second feature of all samples.

The variances and covariance can be calculated from the following formulas:

$$\sigma_1^2 = \frac{\sum_{i=1}^{n} (x_{1i} - \mu_1)^2}{n}, \sigma_2^2 = \frac{\sum_{i=1}^{n} (x_{2i} - \mu_2)^2}{n}, \ \sigma_{12} = \frac{\sum_{i=1}^{n} (x_{1i} - \mu_1)(x_{2i} - \mu_2)}{n}$$

Where μ_1 is the mean for feature 1, μ_2 is the mean for feature 2,, x_{1i} is the value of the first feature for the ith sample, x_{2i} is the value of the second feature for the ith sample, n is the total number of samples.

Estimated value of parameter Covariance Matrix

For Digit 0 training set

The estimated value of the covariance matrix = [[115.28525227 106.23944011] [106.23944011 101.49732487]]

For Digit 1 training set

The estimated value of the covariance matrix = [[31.4515248 50.25141216] [50.25141216 82.70289796]]

Estimated Normal Distributions

The Normal distribution, also known as the Gaussian Distribution, is a continuous probability distribution that has a symmetric bell shaped curve.

The parameters for the Gaussian distribution are the mean (μ) and standard deviation (σ). It is represented as N(μ, σ)

For the specific datasets that were being analyzed, the samples were represented by a 2-D features vector that was drawn from the 2-D normal distribution, therefore a 2-D Gaussian Distribution equation was used:

$$p(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} exp[-\frac{1}{2}(x - \mu)^t \Sigma^{-1}(x - \mu)]$$

d represents the number of features for each sample, which was 2 in our case. x represents the feature vector, containing the values for the features for each sample in the dataset. Σ is the covariance matrix of the distribution, which can be represented as $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. $|\Sigma|$ is the determinant of the matrix, which will provide a scalar value that gives information on the matrix. Σ^{-1} is the inverse of the matrix, which measures the distance from the mean in the space where the covariance structure is normalized. μ represents the mean vector whose components are the means of both the features of the corresponding class.

This expression was used later on when we created a Naive Bayes classifier to predict the class labels of the testing set. It was used to calculate the likelihoods of each sample belonging to a given class, which after multiplying with the prior probabilities, helped provide the posterior probabilities for both the classes. This helped determine the predicted class label.

The classification process and the use of the distributions are further elaborated in the following section.

Classification Method

For this project, the aim was to classify the samples from the given test data (i.e. predict whether the sample image was a 0 or a 1) by implementing a Naive Bayes classifier.

The Naive Bayes classifier predicts categorical outcomes based on Bayes Theorem.

The following is the form for Bayes theorem =

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

Where P(Y|X) is the Posterior Probability for Y given X, P(X|Y) is the likelihood for X given Y, P(Y) and P(X) are the prior probabilities for Y and X respectively.

Since we assume that the two features that we derived from the image (The mean and standard deviations for the brightness of pixels) are independent, then Naive Bayes theorem will be applicable, due to all the features being independent of each other. This is of the form:

$$P(Y|X_1, X_2, X_3, ..., X_n) \propto P(Y) \Pi_{i=1}^{n} p(X_i|Y)$$

Where Y is the class or label, X_1, X_2, … , X_n are the features.

$P(Y|X_1, .., X_n)$ is the posterior probability of Y, which is the probability of the class being Y for the sample given the features X_1 through X_n. Since the features are independent, we can get the posterior probability by multiplying the prior probability of Y (i.e. P(Y)) by the products of all the likelihoods of each feature.

As per the Naive Bayesian classifier, the predicted class of for a given sample is predicted to be the one which has the maximum posterior probability, given all the features. This is of the form:

$$\widehat{Y} = argmax_y P(Y) \Pi_{i=1}^{d} P(X_i|Y)$$

Where D is the total number of features for the data, X is a feature of the dataset, Y is the current label we are considered, and $\hat{Y}$ represents the predicted label, i.e. the one with the highest posterior probability.

We had initially used the corresponding training data to estimate the parameters for the 2D normal distribution for each digit.

To test the classifier, we make predictions on each of the given testing data samples using the above equation, and also make use of the estimated distributions for the respective digit that we had done on the training sets.

We can calculate the prior probabilities of the classes by the following formula:

$$P(Y) = \frac{Total\ number\ of\ samples\ with\ label\ Y\ in\ the\ dataset}{Total\ number\ of\ samples\ in\ the\ dataset}$$

There are just two classes for the digits (0 and 1), and both will have a prior probability of around 0.5 each.

It was given that each of the samples, represented by 2-D feature vectors, is drawn from a 2-D normal distribution, and the values were continuous. Therefore, we can carry out Gaussian Naive Bayes classification for predicting the labels.

Therefore, to calculate the likelihood P(X|Y), where X is any one of the features and Y the given label, we need to use a 2-D Gaussian distribution, which was mentioned in the Estimated Normal Distributions section of this report.

The mean and the covariance matrix used in the calculation in the formula are the ones estimated as the parameters of the normal distributions for that particular digit from the training samples earlier.

After using this method to calculate the posterior probabilities for both the classes, the class with the highest posterior probability is noted down as the prediction for the digit for that sample.

After this has been carried out for all the testing samples, the accuracy is tested.

Final Accuracy

The final classification accuracy of the Naive Bayes classifier that we made was tested by comparing the predictions it made for the labels for the testing data with the actual target labels for these data.

The classification accuracy was calculated separately for digits 0 and 1 as instructed, using their respective given test sets.

The formula used to calculate the accuracy from the test set for each digit was:

$$Classification\ Accuracy\ = \frac{No.\ of\ test\ samples\ with\ equal\ prediction\ as\ the\ target\ value}{Total\ number\ of\ samples\ in\ the\ test\ set\ of\ the\ digit}$$

The classification accuracy for digit 0 was: 97.45%

The classification accuracy for digit 1 was: 96.39%

The classification accuracy was quite high for both digit 0 and digit 1 which proves that the Naive Bayes classifier we implemented works quite reliably.