SYDE 572: Intro to Pattern Recognition
Assigment 2

Cooper Ang (20768006)

October 21, 2022

1 Exercise 1

1.1 Question 1

(14 pts) Assume that probability distribution of your dataset is a multi-variate Gaussian Distribution. Using maximum likelihood estimation (MLE), derive the formulae for estimating the two parameters of the Gaussian distribution i.e. mean and the covariance matrix. Make sure that this is a general expression that can be applied to any d dimensional data.

The maximum likelihood estimation (MLE) with the assumption of a multi-variate Gaussian Distribution treats each parameter as being fixed but unknown. Without knowing the PDF, I can estimate the parameters of the PDF by maximizing the probability that the samples I have came from the PDF with the estimated parameters.

The multi-variate normal distribution is characterized by by the mean vector $\underline{\mu}$ and the covariance matrix Σ:

$$\mathcal{N}_p(\underline{\mu}, \Sigma)$$

Assuming that each sample is independent from each other, to maximize the total likelihood using the estimation of the parameters the sample set probability is the product of the individual sample probabilities:

$$p\left(\{\underline{x}_i\} \mid \underline{\mu}, \Sigma\right) = p\left(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N \mid \underline{\mu}, \Sigma\right) = \prod_{i=1}^{N} p\left(\underline{x}_i \mid \underline{\mu}, \Sigma\right) \tag{1}$$

Subbing in the definition of the multivariate Gaussian distribution:

$$p(\underline{x} \mid \underline{\mu}, \Sigma) = \frac{\left|\Sigma^{-1}\right|^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{x} - \underline{\mu}\right)\right] \tag{2}$$

Taking the log of Equation 2 gives:

$$l(\underline{\mu}, \Sigma) = \sum_{i=1}^{N} \frac{1}{2}\log\left|\Sigma^{-1}\right| - \frac{n}{2}\log 2\pi - \frac{1}{2}\left(\underline{x}_i - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{x}_i - \underline{\mu}\right) \tag{3}$$

To find the mean ($\underline{\mu}$) that maximizes the log likelihood, I can take the derivative with respect to $\underline{\mu}$ and set that equal to zero.

$$\frac{\partial}{\partial \underline{\mu}} l(\underline{\mu}, \Sigma) = \frac{\partial}{\partial \underline{\mu}}\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\underline{x}_i - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{x}_i - \underline{\mu}\right)\right] = 0 \tag{4}$$

$$\sum_{i=1}^{N}\left(\underline{x}_i - \underline{\mu}\right) = 0 \tag{5}$$

$$\sum_{i=1}^{N} (\underline{x}_i) = N\underline{\mu} \tag{6}$$

$$\underline{\mu} = \frac{1}{N} \sum_{i=1}^{N} (\underline{x}_i) \tag{7}$$

So it seems like the $\underline{\mu}$ that provides the maximum likelihood is just the sample mean.

To find the covariance matrix Σ that maximizes the log likelihood, I can take the derivative with respect to Σ^{-1} and set that equal to zero.

$$\frac{\partial l(\underline{\mu}, \Sigma)}{\partial \Sigma^{-1}} = \frac{1}{2} \sum_{i=1}^{N} \frac{\partial \log \left|\Sigma^{-1}\right|}{\partial \Sigma^{-1}} - \frac{1}{2} \sum_{i=1}^{N} \frac{\partial}{\partial \Sigma^{-1}} \left[\left(\underline{x}_i - \underline{\mu}\right)^T \Sigma^{-1} \left(\underline{x}_i - \underline{\mu}\right) \right] \tag{8}$$

$$\frac{\partial l(\underline{\mu}, \Sigma)}{\partial \Sigma^{-1}} = \frac{1}{2} \sum_{i=1}^{N} \frac{\operatorname{cof} \Sigma^{-1}}{|\Sigma^{-1}|} - \frac{1}{2} \sum_{i=1}^{N} \left(\underline{x}_i - \underline{\mu}\right) \left(\underline{x}_i - \underline{\mu}\right)^T \tag{9}$$

$$\frac{\partial l(\underline{\mu}, \Sigma)}{\partial \Sigma^{-1}} = \frac{1}{2} \sum_{i=1}^{N} \left[\Sigma^{-1-1}\right]^T - \frac{1}{2} \sum_{i=1}^{N} \left(\underline{x}_i - \underline{\mu}\right) \left(\underline{x}_i - \underline{\mu}\right)^T \tag{10}$$

$$\frac{1}{2} \sum_{i=1}^{N} [\Sigma]^T - \frac{1}{2} \sum_{i=1}^{N} \left(\underline{x}_i - \underline{\mu}\right) \left(\underline{x}_i - \underline{\mu}\right)^T = 0 \tag{11}$$

$$\sum_{i=1}^{N} [\Sigma]^T = \sum_{i=1}^{N} \left(\underline{x}_i - \underline{\mu}\right) \left(\underline{x}_i - \underline{\mu}\right)^T \tag{12}$$

$$[\Sigma]^T = \frac{1}{N} \sum_{i=1}^{N} \left(\underline{x}_i - \underline{\mu}\right) \left(\underline{x}_i - \underline{\mu}\right)^T \tag{13}$$

Because Σ is a symmetric matrix:

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} \left(\underline{x}_i - \underline{\mu}\right) \left(\underline{x}_i - \underline{\mu}\right)^T \tag{14}$$

The Σ that provides the maximum likelihood is the sample covariance matrix.

1.2 Question 2

(2 pts) Using the derived expressions for the two parameters of the Gaussian distribution, find the probability distribution of the MNIST dataset. Note that you will have to estimate the distributions of the two classes separately. So, you will find two mean vectors and two covariance matrices.

The parameters for the MLE assuming a multivariate Gaussian distribution have the following covariance matrices (Σ) and mean vectors ($\underline{\mu}$). For the full distribution I can sub both parameters into the definition of the multivariate Gaussian distribution in Equation 2.

For class 0:

$$\Sigma = \begin{bmatrix} 215286.02496249 & -23351.22648952 \\ -23351.22648952 & 190498.54227856 \end{bmatrix}$$

$$\underline{\mu} = \begin{bmatrix} 1023.98291519 \\ 17.2161332 \end{bmatrix}$$

For class 1:

$$\Sigma = \begin{bmatrix} 22049.68710465 & -8583.77386087 \\ -8583.77386087 & 380576.59122214 \end{bmatrix}$$

$$\underline{\mu} = \begin{bmatrix} -899.59222882 \\ -15.12476371 \end{bmatrix}$$

1.3 Question 3

(5 pts) Use the maximum likelihood (ML) classifier on the estimated class distributions $p(X|C)$, to make predictions on your test set. Find the prediction error for your ML classifier.

The classifier is applied by taking the maximum likelihood across both PDFs defined by their respective multi-variate Gaussian distributions. Please see code for a detailed walk-through on how that was done.

The prediction error on the test set using the ML classifier is 0.0037825

1.4 Question 4

(3 pts) Plot the decision boundary for your classifier. Note, you don't have to derive the boundary mathematically.

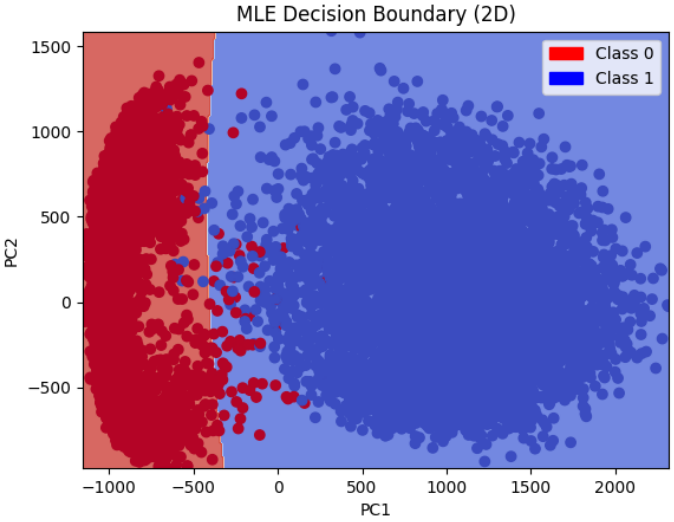

Figure 1: MLE Decision Boundary 2D

1.5 Question 5

(1 pt) Now, let's find the prior probabilities for the two classes ($p(C_1)$ and $p(C_2)$) using the class statistics: $p(C_k) = N_k/N$, where N_k is the number of training data points for class C_k, and N is the total number of training data points.

$$p(C_0) = \frac{5923}{12665} = 0.46766$$

$$p(C_1) = \frac{6742}{12665} = 0.53233$$

1.6 Question 6

(10 pts) Using the prior probabilities, and the estimated class distributions $p(X|C)$, consider the MAP classifier. Make predictions using the MAP classifier on the test set. How does this compare to the ML classifier? Explain.

The prediction error on the test set using the MAP classifier is 0.0037825

Comparing the ML classifier with the MAP classifier shows that the error is the exact same. This would be because the ML classifier assumes that the prior probabilities are equal, where $p(C_0) = 0.5$ and $p(C_1) = 0.5$. Based on the class statistics the prior probabilities are very similar with only a difference of around a few percentage points of probability. It seems like that was not enough to vary the results of the classification when taking into account the prior probability by adding the log prior of the class to the log likelihood of the class.

1.7 Question 7

(5 pts) Compare the results for your MAP and ML classifiers with MED and GED classifiers from the previous assignment. Explain.

The results of the MAP and ML classifier in addition to a reminder from the last assignment of the performance of the MED and GED classifier using the 2D data were as follows (ordered from lowest error to highest):

GED classifier has an error of 0.00331
ML classifier has an error of 0.003782
MAP classifier has an error of 0.003782
MED classifier has an error of 0.005674

From a performance standpoint comparing the results the GED classifier performs the best with the ML and MAP classifier coming in as a tie for second place, and the MED classifier performing the worst. Although the overall performance of all the classifiers is relatively good. The MED classifier probably performs the worst as it does not take into account the class statistics and only creates a linear decision boundary. The ML, MAP, and GED classifiers all take into account the class statistics when creating the classifier thus they all are slightly better, with a more complex decision boundary to correct for the class statistics.

2 Exercise 2

2.1 Question 1

(10 pts) Let's assume that the probability distribution of your dataset is an exponential distribution (Wiki). In this case, you only need to estimate a single parameter λ. Use MLE to derive an expression for finding λ.

The PDF of the exponential distribution is defined as:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{15}$$

$$\mathcal{L}(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x} = \lambda^n \exp\left[-\lambda \sum_{i=1}^{n} x_i\right] \tag{16}$$

Taking the log of Equation 16 gives:

$$l\left(\lambda; x_1, \ldots, x_n\right) = n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i \tag{17}$$

To find λ to maximize the log-likelihood I can take the derivative with respect to λ and set that to zero:

$$\frac{d}{d\lambda} l\left(\lambda; x_1, \ldots, x_n\right) = \frac{d}{d\lambda}\left(n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i\right) \tag{18}$$

$$= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \tag{19}$$

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i} \tag{20}$$

The λ that maximizes the likelihood is the reciprocal of the sample mean.

2.2 Question 2

(10 pts) Now, let's assume the probability distribution of your dataset is uniform (Zhang Ch. 19.8). In this case, you will estimate two parameters a and b. Use MLE to derive an expression for finding the two parameters.

The PDF of the uniform distribution is defined as:

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \tag{21}$$

$$\mathcal{L}(a, b; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; a, b) = \frac{1}{(b-a)^n} \tag{22}$$

Taking the log of Equation 22 gives:

$$l\,(a, b; x_1, \ldots, x_n) = -n \log(b - a) \tag{23}$$

To find a to maximize the log-likelihood I can take the derivative with respect to a:

$$\frac{\partial}{\partial a} l\,(a, b; x_1, \ldots, x_n) = \frac{n}{(b-a)} \tag{24}$$

Since the derivative with respect to a is monotonically increasing for the maximum likelihood the best estimate of a would be the largest a possible, which is:

$$a = min(x_1, \ldots, x_n) \tag{25}$$

To find b to maximize the log-likelihood I can take the derivative with respect to b and set that to zero:

$$\frac{\partial}{\partial b} l\,(a, b; x_1, \ldots, x_n) = -\frac{n}{(b-a)} \tag{26}$$

Since the derivative with respect to b is monotonically decreasing for the maximum likelihood the best estimate of b would be the smallest b possible, which is:

$$b = max(x_1, \ldots, x_n) \tag{27}$$

2.3 Question 3

(15 pts) Use the ML classifier on the estimated class distributions from exponential estimation, uniform distribution, and Gaussian distribution (you will need to recompute the mean and covariance matrices for 1-d data), to make predictions on the test set. Find the prediction error for the three different ML classifiers (for the three estimated probability distributions).

Error for MLE Exponential Classifier: 0.0264775
Error for MLE Uniform Classifier: 0.02316784
Error for MLE Gaussian Classifier: 0.003782

2.4 Question 4

(5 pts) Which of the estimated probability distributions best represent your data? Explain.

In theory the estimated probability distribution with the lowest error should represent the data the best. This is because running the test data through the classifier will result in a classification based on the estimates of the probability distribution and a better estimate would result in a better generalization to the pattern recognition task. In this case it would be the Gaussian distribution. In addition if you take a look at the histogram of the 1-D MNIST data in Figure 2, you can clearly eyeball that the distribution looks most like a normal distribution, and not like an exponential or uniform distribution.

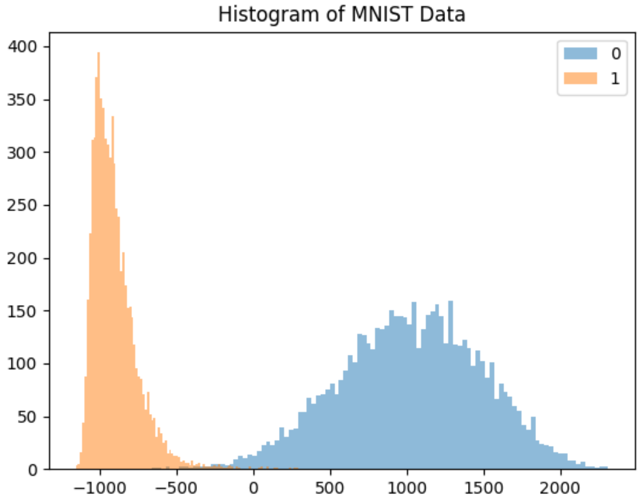

Figure 2: Histograms of MNIST Data

3 Exercise 3

3.1 Question 1

(20 pts) Let's assume the probability distribution of your dataset is not just a single Gaussian distribution but a sum of two Gaussian distributions:

$$p(X|\theta) = \frac{1}{2}[\mathcal{N}(X|\theta_1, \theta_2) \times \mathcal{N}(X|\theta_3, \theta_4)] \tag{28}$$

The only difference between this expression and the one in your notes for parametric estimation, is that instead of estimating the mean and variance for one Gaussian, you will be estimating the mean and variances of two Gaussian distributions. Using MLE, derive an expression for finding the four parameters of your probability distribution defined by eq. 28.

$$p(X|\theta) = \frac{1}{2}[\mathcal{N}(X|\theta_1, \theta_2) \times \mathcal{N}(X|\theta_3, \theta_4)] \tag{29}$$

$$L(\theta; x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{2}\left[\frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{(x_i - \theta_1)^2}{2\theta_2}\right] \times \frac{1}{\sqrt{2\pi\theta_4}} \exp\left[-\frac{(x_i - \theta_3)^2}{2\theta_4}\right]\right] \tag{30}$$

$$L(\theta; x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{2} \frac{1}{\sqrt{2\pi\theta_2}} \frac{1}{\sqrt{2\pi\theta_4}} \exp\left[-\frac{(x_i - \theta_1)^2}{2\theta_2} - \frac{(x_i - \theta_3)^2}{2\theta_4}\right] \tag{31}$$

Taking the log of Equation 31 gives:

$$l\left(\theta; x_1, \ldots, x_N\right) = \sum_{i=1}^{N} -\frac{1}{2}\log(8\pi\theta_2\theta_4) - \frac{\left(x_i - \theta_1\right)^2}{2\theta_2} - \frac{\left(x_i - \theta_3\right)^2}{2\theta_4} \tag{32}$$

To find θ_1 to maximize the log-likelihood I can take the derivative with respect to θ_1 and set that to zero:

$$\frac{\partial}{\partial\theta_1} l\left(\theta; x_1, \ldots, x_N\right) = \sum_{i=1}^{N} \frac{\left(x_i - \theta_1\right)}{\theta_2} = 0 \tag{33}$$

$$\sum_{i=1}^{N} \left(x_i - \theta_1\right) = 0 \tag{34}$$

$$\theta_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{35}$$

To find θ_3 to maximize the log-likelihood I can take the derivative with respect to θ_3 and set that to zero:

$$\frac{\partial}{\partial\theta_3} l\left(\theta; x_1, \ldots, x_N\right) = \sum_{i=1}^{N} \frac{\left(x_i - \theta_3\right)}{\theta_4} = 0 \tag{36}$$

$$\sum_{i=1}^{N} \left(x_i - \theta_3\right) = 0 \tag{37}$$

$$\theta_3 = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{38}$$

So it seems both θ_1 and θ_3 are the sample means.

To find θ_2 to maximize the log-likelihood I can take the derivative with respect to θ_2 and set that to zero:

$$\frac{\partial}{\partial\theta_2} l\left(\theta; x_1, \ldots, x_N\right) = \frac{\partial}{\partial\theta_2}\sum_{i=1}^{N} -\frac{1}{2}\log(8\pi\theta_2\theta_4) + \frac{\partial}{\partial\theta_2}\sum_{i=1}^{N} -\frac{\left(x_i - \theta_1\right)^2}{2\theta_2} = 0 \tag{39}$$

$$-\frac{N}{2\theta_2} + \frac{1}{2}\sum_{i=1}^{N} \frac{\left(x_i - \theta_1\right)^2}{\theta_2^2} = 0 \tag{40}$$

$$\theta_2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \theta_1\right)^2 \tag{41}$$

To find θ_4 to maximize the log-likelihood I can take the derivative with respect to θ_4 and set that to zero:

$$\frac{\partial}{\partial\theta_4} l\left(\theta; x_1, \ldots, x_N\right) = \frac{\partial}{\partial\theta_4}\sum_{i=1}^{N} -\frac{1}{2}\log(8\pi\theta_2\theta_4) + \frac{\partial}{\partial\theta_4}\sum_{i=1}^{N} -\frac{\left(x_i - \theta_3\right)^2}{2\theta_4} = 0 \tag{42}$$

$$-\frac{N}{2\theta_4} + \frac{1}{2}\sum_{i=1}^{N} \frac{\left(x_i - \theta_3\right)^2}{\theta_4^2} = 0 \tag{43}$$

$$\theta_4 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \theta_3\right)^2 \tag{44}$$

Since both θ_1 and θ_3 are the sample means, substituting into θ_2 and θ_4, results in θ_2 and θ_4 just being the sample variance.