UiA Fakultet for teknologi og realfag

MA-447-G 24H AI Mathematics

Assignments

Autumn 2024

Andrea Zatti

7. november 2024

Mandatory Group Declaration

Each student is solely responsible for familiarizing themselves with the legal aids, guidelines for their use, and rules regarding source usage. The declaration aims to raise awareness among students of their responsibilities and the consequences of cheating. Lack of declaration does not exempt students from their responsibilities.

1.	We hereby declare that our submission is our own work and that we have not used other sources or received any help other than what is mentioned in the submission.	Yes / No
2.	**We further declare that this submission:** • Has not been used for any other examination at another department/university/college domestically or abroad. • Does not reference others' work without it being indicated. • Does not reference our own previous work without it being indicated. • Has all references included in the bibliography. • Is not a copy, duplicate, or transcription of others' work or submission.	Yes / No
3.	We are aware that violations of the above are considered to be cheating and can result in cancellation of the examination and exclusion from universities and colleges in Norway, according to the Universities and Colleges Act, sections 4-7 and 4-8 and the Examination Regulation, sections 31.	Yes / No
4.	We are aware that all submitted assignments may be subjected to plagiarism checks.	Yes / No
5.	We are aware that the University of Agder will handle all cases where there is suspicion of cheating according to the university's guidelines for handling cheating cases.	Yes / No
6.	We have familiarized ourselves with the rules and guidelines for using sources and references on the library's website.	Yes / No
7.	We have in the majority agreed that the effort within the group is notably different and therefore wish to be evaluated individually. Ordinarily, all participants in the project are evaluated collectively.	Yes / No

Publishing Agreement

Authorization for Electronic Publication of Work The author(s) hold the copyright to the work. This means, among other things, the exclusive right to make the work available to the public (Copyright Act. §2).

Theses that are exempt from public access or confidential will not be published.

We hereby grant the University of Agder a royalty-free right to make the work available for electronic publication:	Yes / No
Is the work confidential?	Yes / No
Is the work exempt from public access?	Yes / No

Innhold

1 Introduction

Gradient Descent is one of the most widely used optimization algorithms in the field of artificial intelligence, applied to minimize loss functions and find the optimal parameters of a model. However, over the years, numerous variants of the algorithm have been proposed to improve convergence speed, accuracy, and computational efficiency. For the preparation of this report, the papers that was consulted: Survey of Gradient Descent Variants and Evaluation Criteria[1], as it proposes evaluations of the most popular variants of the Gradient Descent algorithm.

1.1 Gradient Descent

The objective of the Gradient Descent algorithm is to find the optimal values of a model's parameters by minimizing a cost function or loss function. This process is iterative and relies on calculating the gradient of the function with respect to the parameters. Specifically, these elements are:

- **Cost Function** $L(\theta)$: In machine learning, the goal is to minimize a cost function $L(\theta)$, where θ represents the model's parameters. This function measures the model's error on a dataset.
- **Gradient** $\nabla L(\theta)$: The gradient is a vector of partial derivatives of the cost function with respect to the model's parameters. It indicates the direction of the steepest change in the function, i.e., the direction in which the function increases most rapidly.
- **Learning Rate** γ: The learning rate determines the size of the step the algorithm takes in the direction of the gradient. If γ is too large, the algorithm may overshoot the minimum and fail to converge; if it is too small, the algorithm may converge very slowly.

The Gradient Descent algorithm updates the parameters θ at each iteration according to the following formula:

$$\theta = \theta - \gamma \cdot \nabla L(\theta)$$

Where:

- θ is the parameter vector,
- γ is the learning rate,
- $\nabla L(\theta)$ is the gradient of the cost function with respect to θ.

1.2 Variants

The paper describes several variants of the Gradient Descent algorithm, each addressing the limitations of the standard version:

- **Stochastic Gradient Descent (SGD)**: This variant updates the parameters using a single training example per iteration, which makes the algorithm faster but introduces noise into the update process.
- **Mini-Batch Gradient Descent**: It is a compromise between batch gradient descent (which uses the entire dataset) and SGD, and uses small batches of data for each update, improving stability without sacrificing too much speed.
- **Momentum-Based Gradient Descent**: Introduces a momentum term to accelerate convergence and reduce oscillations in problems with non-linear functions.
- **Adam (Adaptive Moment Estimation)**: Combines the techniques of AdaGrad and RMSProp to dynamically adjust the learning rate during optimization.
- **Adagrad (Adaptive Gradient Algorithm)**: It adjusts the learning rate for each parameter independently, reducing the rate for parameters that receive frequent updates.
- **RMSProp (Root Mean Square Propagation)**: It maintains a constant learning rate by using an exponential moving average of the squared gradients, preventing the rate from becoming too small.

2 Evaluation Criteria

The paper discusses various criteria for evaluating and comparing algorithm variants:

- **Convergence Speed**: Measures how quickly an algorithm reaches the minimum of the objective function.
- **Generalization Performance**: Concerns the algorithm's ability to make accurate predictions on unseen data.
- **Computational Efficiency**: Refers to the computational cost in terms of time and memory.
- **Stability and Robustness**: Assesses the algorithm's resistance to noise and unstructured data.
- **Scalability**: Measures how the algorithm performs on large datasets and complex models.
- **Convergence Point**: Analyzes where the algorithm stabilizes in the optimization landscape.

3 Conclusion

The paper does not specify a single variant of the Gradient Descent algorithm as universally superior. Instead, it emphasizes that the choice of the best variant depends on the specific use case and optimization criteria, such as convergence speed, generalization performance, computational efficiency, and robustness to noise. Different variants, like Adam, RMSProp, Momentum-based Gradient Descent, and others, offer distinct advantages in different contexts, such as training deep neural networks or handling large datasets. In summary, there is no single best variant; the optimal choice depends on the problem's characteristics and the desired evaluation criteria, and this is also what I have tried to show in my Python notebooks with the different functions and the various Gradient Descent variants used.

4 Introduction

For the preparation of this report, the following papers were consulted: "Temperature-Aware Processor Frequency Assignment for MPSoCs Using Convex Optimization"[2] and "Temperature Control of High-Performance Multi-core Platforms Using Convex Optimization"[3], which were published one year apart by the same authors. The two papers propose a solution to the problem of thermal optimization in multi-processor systems based on convex optimization to regulate the operating frequency of individual cores.

5 Problem Description

The papers address a critical problem in thermal and performance optimization for multi-processor systems-on-chip (MPSoC). Specifically, as the computing capabilities of MPSoC systems increase, power dissipation also rises, leading to a significant increase in chip temperature. This thermal rise can cause timing delays, compromise system safety, and even permanently damage the devices. One of the main challenges is optimizing the operating frequencies of the processors so that the system operates with maximum efficiency without exceeding the imposed temperature and power limits to avoid damage. Traditional methods such as Dynamic Frequency Scaling (DFS), although effective in reducing power consumption and managing temperature, have some key limitations:

- **Reactivity**: The system only reacts when the temperature exceeds a critical threshold, meaning that cores can operate for a certain period at temperatures higher than the allowed maximum.
- **Limited scalability**: Frequency management is often performed independently for each core, without considering the thermal impact on neighboring cores.
- **Hotspots**: DFS is not effective in reducing thermal gradients between cores.

6 Problem Solution

The authors developed a new solution to address this problem, called the Pro-Temp method, which is structured in two main phases:

1. **Off-line phase** (design-time): Optimal frequencies for each core are calculated to meet thermal and workload constraints. This is done using a thermal model of the chip and a convex optimization approach.
2. **On-line phase** (run-time): During system execution, a thermal management unit monitors core temperatures and applies DFS based on pre-calculated information.

The thermal model used is represented by heat equations that describe how the temperature of each core varies over time as a function of power consumption and the temperature of neighboring cores. The authors use a convex model to find the optimal operating frequencies that respect the temperature constraints and minimize power consumption. The optimization objective function is:

$$\min \sum_i p_i \tag{1}$$

subject to:

$$t_{k+1,i} = t_{k,i} + \sum_{j \in \text{Adji}} a_{i,j}(t_{k,j} - t_{k,i}) + b_i p_i \tag{2}$$

where p_i is the power consumption of each core, $t_{k,i}$ is the temperature of core i at time k, and parameters $a_{i,j}$ and b_i describe the thermal behavior of the chip[3]. For a more detailed view, the complete model is presented below:

$$\begin{aligned}
\text{min:} \quad & \sum_{i=1}^{n} p_i && (3)\\
\text{s.t.} \quad & t_{0,i} = t_{\text{start}}, \quad \forall i && (4)\\
& t_{k+1,i} = t_{k,i} + \sum_{\forall j \in \text{Adj}_i} a_{i,j}(t_{k,j} - t_{k,i}) + b_i p_i, \quad \forall i, k && (5)\\
& t_{k,i} \leq t_{\max}, \quad \forall k, i && (6)\\
& p_{\max} \frac{f_i^2}{f_{\max}^2} \leq p_{i,k}, \quad \forall i, k && (7)\\
& \sum_{i=1}^{n} f_i \geq n \times f_{\text{target}} && (8)\\
& f_i \geq 0, \quad \forall i && (9)
\end{aligned}$$

The authors conducted experiments on a multi-core platform based on Sun Microsystems' Niagara architecture. The results demonstrate that the Pro-Temp method consistently maintains core temperatures below the maximum allowed threshold, unlike traditional DFS methods, which frequently result in thermal violations. Furthermore, the findings reveal that Pro-Temp significantly reduces task waiting times and enhances overall system performance. By proactively preventing overheating, the method ensures that cores remain operational at optimal frequencies, leading to improved efficiency and reliability.

7 Conclusion

The paper concludes that the proposed approach, based on convex optimization, significantly improves the thermal management of MPSoCs, ensuring that the system remains operational within the imposed temperature and power limits. Moreover, this approach can be leveraged to explore the design space of MPSoC systems, enabling informed decisions regarding thermal and power configurations during the design phase.

8 Question 1

8.1 Problem 1 a

Find eigenvectors and eigenvalues of: $\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$

$$A - \lambda I = \begin{bmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \end{bmatrix}$$

$$\begin{aligned} \det(A) &= (1-\lambda)(2-\lambda)(1-\lambda) + 0 + 0 - (1-\lambda) - (1-\lambda) = \\ &= (1-\lambda)^2(2-\lambda) - 2(1-\lambda) = \left(1 + \lambda^2 - 2\lambda\right)(2-\lambda) - 2 + 2\lambda = \\ &= 2 - \lambda + 2\lambda^2 - \lambda^3 - 4\lambda + 2\lambda^2 - 2 + 2\lambda = -\lambda^3 + 4\lambda^2 - 3\lambda = \\ &= -\lambda\left(\lambda^2 - 4\lambda + 3\right) = -\lambda(\lambda-1)(\lambda-3) \end{aligned}$$

$-\lambda_1\left(\lambda_2 - 1\right)\left(\lambda_3 - 3\right) = 0 \quad \lambda_1 = 0 \quad \lambda_2 = 1 \quad \lambda_3 = 3$

$$\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}_{\lambda_1=0} \quad \begin{cases} x - y = 0 \\ -x + 2y - z = 0 \\ -y + z = 0 \end{cases}$$

$$v_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad \lambda_1 = 0$$

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix}_{\lambda_2=1} \quad \begin{cases} -y = 0 \\ -x + y - z = 0 \\ -y = 0 \end{cases}$$

$$v_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \qquad \lambda_2 = 1$$

$$\begin{bmatrix} -2 & -1 & 0 \\ -1 & -1 & -1 \\ 0 & -1 & -2 \end{bmatrix}_{\lambda_3=3} \quad \begin{cases} -2x - y = 0 \\ -x - y - z = 0 \\ -y - 2z = 0 \end{cases}$$

$$v_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \qquad \lambda_3 = 3$$

$Trace(A) = \sum_{i=1}^{3} \lambda_i = 4$

8.2 Problem 1 b

Verify from the following matrix that sum of eigenvalues is equal to the trace of the matrix:

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} 2-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 2-\lambda \end{bmatrix}$$

$$\begin{aligned} \det(A) &= (2-\lambda)^3 + 0 + 0 + 0 - (2-\lambda) - (2-\lambda) = \\ &= 8 - \lambda^3 - 12\lambda + 6\lambda^2 - 4 + 2\lambda = \\ &= -\lambda^3 + 6\lambda^2 - 10\lambda + 4 = -(\lambda - 2)\left(\lambda^2 - 4\lambda + 2\right) \end{aligned}$$

$$\begin{array}{l} (2-\lambda)\left(\lambda^2 - 4\lambda + 2\right) = 0 \quad \lambda_1 = 2 \quad \lambda_2 = 2 + \sqrt{2} \quad \lambda_3 = 2 - \sqrt{2} \\ \lambda_{2,3} = 2 \pm \sqrt{4-2} \end{array}$$

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}_{\lambda_1 = 2} \begin{cases} -y = 0 \\ -x - z = 0 \\ -y = 0 \end{cases}$$

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad \lambda_1 = 2$$

$$\begin{bmatrix} -\sqrt{2} & -1 & 0 \\ -1 & -\sqrt{2} & -1 \\ 0 & -1 & -\sqrt{2} \end{bmatrix}_{\lambda_2 = 2+\sqrt{2}} \begin{cases} -\sqrt{2}x - y = 0 \\ -x - \sqrt{2}y - z = 0 \\ -y - \sqrt{2}z = 0 \end{cases}$$

$$v_2 = \begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \end{bmatrix} \quad \lambda_2 = 2 + \sqrt{2}$$

$$\begin{bmatrix} +\sqrt{2} & -1 & 0 \\ -1 & +\sqrt{2} & -1 \\ 0 & -1 & +\sqrt{2} \end{bmatrix}_{\lambda_3 = 2-\sqrt{2}} \begin{cases} +\sqrt{2}x - y = 0 \\ -x + \sqrt{2}y - z = 0 \\ -y + \sqrt{2}z = 0 \end{cases}$$

$$v_3 = \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix} \quad \lambda_3 = 2 - \sqrt{2}$$

$$Trace(A) = 6 = \sum_{i=1}^{3} \lambda_i = \lambda_1 + \lambda_2 + \lambda_3 = 2 + 2 + \sqrt{2} + 2 - \sqrt{2} = 6$$

8.3 Problem 1 c

Verify from the above matrix that product of eigenvalues is equal to the determinant of the matrix.

$\det(A) = 8 + 0 + 0 + 0 - 2 - 2 = 4$

$\prod_{i=1}^{3} \lambda_i = 2(2 + \sqrt{2})(2 - \sqrt{2}) = 2(4 - 2) = 4 = \det(A)$

8.4 Note

At the end of the document, the pages with handwritten exercises are attached, in case any steps are not sufficiently clear in the LaTeX transcription.

9 Question 4

9.1 Introduction

In the field of statistics and research, data loss occurs when some of the observations expected in a study are unavailable or omitted. This issue can arise from measurement errors, non responses from participants, or technical problems during data collection. Data loss can compromise the accuracy of results, reduce the statistical power of analyses, and introduce bias, thus affecting the reliability of the conclusions drawn. Therefore, it is crucial to implement data management and imputation techniques to mitigate these negative effects. For the preparation of this report, the following papers were consulted: ”Review of the Methods for Handling Missing Data in Longitudinal Data Analysis”[4] and ”The prevention and handling of the missing data”[5].

Hyun Kang's first paper focuses on the importance of handling missing data in medical research, particularly in clinical trials. The paper provides an overview of the various types of missing data, discusses techniques to address them, and offers recommendations for effective data management.

The second paper by Michikazu Nakai and Weiming Ke discusses various methods for handling missing data in longitudinal data analysis. Longitudinal analysis involves studies where subjects are repeatedly observed and measured over time. Missing data is a common issue in such studies, often resulting from individuals dropping out before the study is completed.

9.2 Problem Solution

Both papers distinguish three fundamental types of missing data, based on Rubin's classification:

- **MCAR (Missing Completely At Random)**: Missing data are neither related to the observed values nor to the missing ones. This is the strongest and least common assumption.
- **MAR (Missing At Random)**: Missing data are related to the observed values but not to the missing ones. This is a more realistic assumption for many studies.
- **MNAR (Missing Not At Random)**: Missing data depend directly on the missing values themselves. This type of missingness is more complex to handle and requires specific models for proper treatment.

Both papers mention numerous techniques for addressing issues related to data loss, some of which are simpler and easier to implement but introduce significant bias and reduce statistical power. Other methods, while more advanced, offer greater accuracy and robustness but require more computational resources and expertise for proper implementation. For the sake of simplicity, I will not cover all the methods discussed, as they are numerous, but will focus on the two most advanced techniques: Expectation Maximization and Multiple Imputation.

9.2.1 Expectation Maximization

The Expectation Maximization (EM) algorithm is a maximum likelihood method used to create a new dataset where missing values are imputed using estimates derived from maximum likelihood techniques. The process begins with the expectation step, during which variables (such as variances, covariances, and means) are estimated, possibly using listwise deletion. These estimates are then used to create a regression equation to predict the missing data. In the maximization step, these equations are applied to fill in the missing values. The expectation step is then repeated with the updated parameters, and this cycle continues until the system stabilizes when the covariance matrix of successive iterations becomes virtually identical. A key feature of EM imputation is that when the complete dataset is generated, a random disturbance term is incorporated into each imputed value to reflect the uncertainty associated with the imputation. However, EM has some drawbacks: it can take a long time to converge, especially when there is a high percentage of missing data, and it may be too complex for some statisticians to adopt. Additionally, this approach can lead to biased parameter estimates and underestimation of the standard error. In the EM imputation method, a predicted value based on the available variables for each case is substituted for the missing data. Since single imputation does not account for the variability between multiple imputations, it tends to underestimate standard errors and overestimate the level of precision.

9.2.2 Multiple Imputation

Multiple Imputation (MI) is the most popular method for handling missing data. It involves replacing each missing item with two or more acceptable values, representing a distribution of possibilities. Once the imputed dataset is generated, analysis can be performed using standard statistical software, making the process simple. MI also produces valid inferences, such as standard errors or p-values, because it accounts for the uncertainty of missing data. MI can be efficient even with a small number of imputations, especially when the variance between imputations is not large. However, MI has some disadvantages. First, imputing values for missing data introduces variability, which may ignore individual variation. Second, the uncertainty inherent in missing data is not fully addressed since the analysis does not distinguish between observed and imputed values. Additionally, MI is more work intensive compared to single imputation, as it requires creating and analyzing multiple imputed datasets.

9.3 Conclusion

Both papers conclude that optimal handling of missing data should begin at the study design phase, aiming to prevent data loss through careful planning and monitoring. However, when missing data is unavoidable, advanced methods such as Multiple Imputation or Expectation Maximization are preferable, as they provide more accurate estimates compared to simpler methods like Listwise Deletion or Last Observation Carried Forward, which can introduce bias and lead to inaccurate results. The choice of the appropriate method depends on the type of missing data (MCAR, MAR, MNAR). For MCAR or MAR data, Multiple Imputation and Expectation Maximization are recommended, while for MNAR data, more complex models like the Selection Model or Pattern Mixture Model are required. Correctly identifying the missing data mechanism is crucial to avoid bias and ensure valid results.

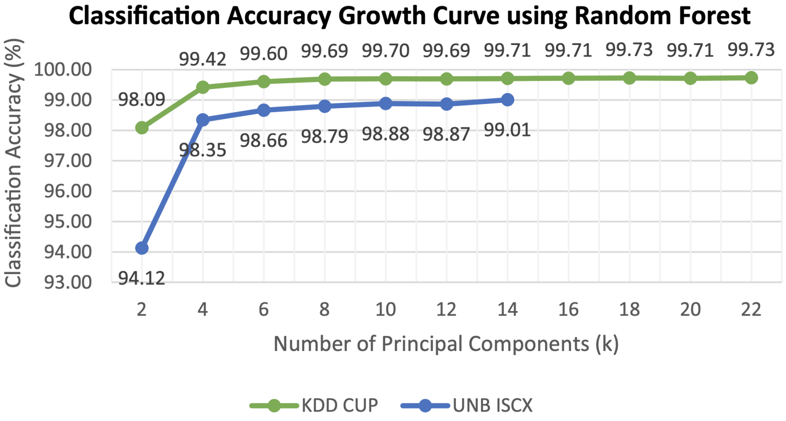
10 Question 3

10.1 Introduction

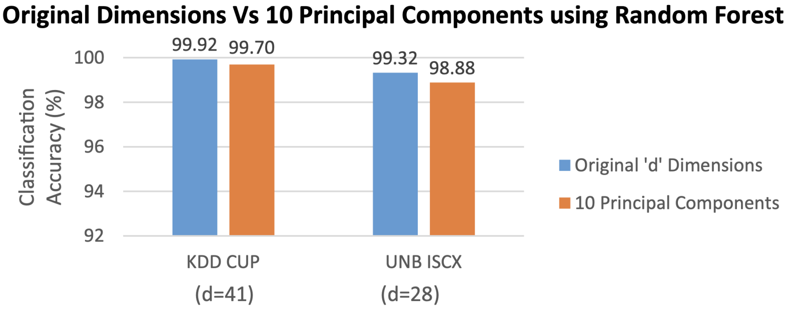
Principal Component Analysis (PCA) is a powerful technique used for dimensionality reduction while preserving the variance in the data. This makes PCA particularly useful for reducing computational time and costs in tasks that are typically resource-intensive. The following two papers address different problems, both of which are resolved through the application of PCA. In the paper "Dimensionality Reduction Using Principal Component Analysis for Network Intrusion Detection"[6] by Keerthi Vasan and B. Surendiran, the authors explore the use of PCA for dimensionality reduction in the context of network intrusion detection. The study focuses on the effectiveness of PCA in reducing the number of features used for analyzing network traffic without compromising classification accuracy. The second paper, "Feature Selection Using Principal Component Analysis"[7] by Fengxi Song, Zhongwei Guo, and Dayong Mei, presents a method for feature selection using PCA. Although PCA is traditionally employed for dimensionality reduction by transforming data into a space of principal components, this study demonstrates that PCA can also be leveraged to select significant features from a dataset.

10.2 Dimensionality Reduction Using Principal Component Analysis for Network Intrusion Detection

Traffic analysis in a network is conducted to understand the nature of the traffic, monitor activities, and manage potential threats in traffic flows. Hostile traffic aimed at a host may either overwhelm the target with a large volume of packets, disrupting its functionality, or probe the target to identify vulnerabilities for a potential attack. These vulnerabilities can then be exploited for a more precise and damaging assault on the target. Relying only on signature-based intrusion detection as a defense method is both insecure and ineffective. However, it is far less computationally demanding, especially compared to machine learning and data mining algorithms. For this reason, using PCA to preprocess the data before applying machine learning algorithms is an excellent strategy. It helps reduce the data's dimensionality and, consequently, lowers the computational cost of the analysis without compromising classification accuracy. The authors conducted experiments on two benchmark datasets: KDD Cup 1999 and UNB ISCX 2012, which are labeled network traffic datasets. The KDD Cup dataset contains 41 features, while UNB ISCX contains 28. The data reduced by PCA was then used with various classifiers, including C4.5 and Random Forest. As we can see in the following figures, the classification accuracy growth curve using Random Forest[10.2] stabilizes at around 8 to 10 principal components and no longer improves significantly. Furthermore, in the second image[10.2], we can see that the performance of Random Forest using the original dimensions or the 10 principal components is very similar. In both cases, this is achieved by using far fewer than half of the original features of the datasets. The authors also define a reduction ratio, which is used to quantify the extent of dimensionality reduction. The reduction ratio for PCA (RRPCA) is the ratio of the number of target dimensions to the number of original dimensions. The lower the RRPCA value, the higher the efficiency of the PCA. Given the values of $k_{ideal} = 10$, $d_{KDD} = 41$ and $d_{ISCX} = 28$, these are the $RRPCA$ values for the respective datasets: $RRPCA_{KDD} = 0.24$ and $RRPCA_{KDD} = 0.36$.



Figur 1: Accuracy performance [6]



Figur 2: Dimensions comparison [6]

10.3 Feature Selection Using Principal Component Analysis

Feature selection is a critical technique in many areas of computational science, including computer vision, pattern recognition, and machine learning. It helps reduce the dimensionality of data and improves computational efficiency without compromising accuracy. Traditional feature selection methods include approaches based on mutual information, feature similarity, and optimization algorithms such as ant colony optimization or genetic algorithms. While PCA typically transforms data into a new space, the authors demonstrate that the eigenvectors of the covariance matrix can be leveraged to evaluate the importance of individual features. The proposed method selects only the most significant features based on these eigenvectors. The process begins with the calculation of the covariance matrix from the original data, followed by the extraction of eigenvalues and eigenvectors. The eigenvectors associated with the largest eigenvalues are used to identify the most relevant features. Features are selected by assessing their contribution to the feature extraction process based on PCA. Those features that contribute less can be eliminated, simplifying the dataset without losing essential information. The method was tested on various facial recognition datasets like the ORL, AR, and Feret databases. The results show that the method significantly reduces the dimensionality of the images (e.g., from 2576 to 1000 dimensions) without compromising recognition accuracy. Two versions of the method were tested: one that uses feature selection followed by classification, and another that also applies PCA after feature selection. In particular, the first method appears to offer slightly better performance when considering the classification error.

10.4 Conclusion

PCA has proven highly effective for both feature selection and dimensionality reduction across various domains, specifically in our cases of network intrusion detection and face recognition. By reducing the number of dimensions, PCA helps preserve high classification accuracy while significantly decreasing computational costs. The experiments presented in both papers consistently show that using a reduced set of principal components achieves performance comparable to the original feature set, underlining PCA's value in optimizing accuracy and efficiency.

11 Question 4

In a city, 47% of the adults are males. One adult is randomly selected for a survey involving credit card usage.

11.1 Problem 4 a

Find the prior probability that the selected person is a female.

$M \to$ event: the selected person is a male

$F \to$ event: the selected person is a female

$\mathbb{P}(M) = 47\% \quad \mathbb{P}(F) = 1 - \mathbb{P}(M) = 53\%$

11.2 Problem 4 b

It is later learned that the selected survey subject was drinking soda. Also, 7.5% of males have soda, whereas 8.2% of females have soda. Use this additional information to find the probability that the selected subject is a male.

$S \to$ event: the selected person drinks soda

$\mathbb{P}(S \mid M) = 7{,}5\% \quad \mathbb{P}(S \mid F) = 8{,}2\%$

$\mathbb{P}(S) = \mathbb{P}(M)\mathbb{P}(S \mid M) + \mathbb{P}(F)\mathbb{P}(S \mid F) = 3{,}525\% + 4{,}346\% = 7{,}871\%$

$$\mathbb{P}(M \mid S) = \frac{\mathbb{P}(S \mid M)\mathbb{P}(H)}{\mathbb{P}(S)} = \frac{(7{,}5\%) \cdot (47\%)}{7{,}871\%} \approx 44{,}785\%$$

12 Question 5

A pile of 8 playing cards has 4 kings, 2 aces and 2 queens. A second pile of 8 playing cards has 1 ace, 4 queens and 3 kings. Choose a card at random from the first pile and place it on the second. Then shuffle the second pile and you choose a card at random from the second pile. If the card drawn from the second deck was an ace, what is the probability that the first card was also an ace?

$A_f \to$ event: ace from the first pile

$A_s \to$ event: ace from the second pile

$$\mathbb{P}(A_f) = \frac{1}{4}$$

$$
\begin{aligned}
\mathbb{P}\left(A_s\right) &= \mathbb{P}\left(A_f\right)\mathbb{P}\left(A_s \mid A_f\right) + \mathbb{P}\left(A_f^c\right)\mathbb{P}\left(A_s \mid A_f^c\right) = \\
&= \frac{1}{4}\cdot\frac{2}{9} + \frac{3}{4}\cdot\frac{1}{9} = \frac{2}{36} + \frac{3}{36} = \frac{5}{36}
\end{aligned}
$$

$$
\mathbb{P}(A_f \mid A_s) = \frac{\mathbb{P}(A_s \mid A_f)\mathbb{P}(A_f)}{\mathbb{P}(A_s)} = \frac{\frac{2}{9}\cdot\frac{1}{4}}{\frac{5}{36}} = \frac{2}{5}
$$

13 Question 6

I have two boxes of melons and lemons. In box 1, there are 15 melons and 16 lemons, in box 2 there are 16 melons and 15 lemons. I randomly pick a box and then in this box randomly pick a fruit. What is the probability that I picked box 2 given that I picked a lemon?

$B_1 \rightarrow$ event: pick the box 1

$B_2 \rightarrow$ event: pick the box 2

$L \rightarrow$ event: pick a lemon

$$
\mathbb{P}(B_1) = \mathbb{P}(B_2) = \frac{1}{2}
$$

$$
\begin{aligned}
\mathbb{P}(L) &= \mathbb{P}\left(B_1\right)\mathbb{P}\left(L \mid B_1\right) + \mathbb{P}\left(B_2\right)\mathbb{P}\left(L \mid B_2\right) = \\
&= \frac{1}{2}\cdot\frac{16}{31} + \frac{1}{2}\cdot\frac{15}{31} = \frac{31}{62} = \frac{1}{2}
\end{aligned}
$$

$$
\mathbb{P}\left(B_2 \mid L\right) = \frac{\mathbb{P}\left(L \mid B_2\right)\mathbb{P}\left(B_2\right)}{\mathbb{P}(L)} = \frac{\frac{15}{31}\cdot\frac{1}{2}}{\frac{1}{2}} = \frac{15}{31}
$$

14 Question 2

The table below shows the number of survey subjects who have received and not received a speeding ticket in the last year, and the color of their car.

	Speeding ticket	No speeding ticket	Total
Red car	15	135	150
Not red car	45	470	515
Total	60	605	665

14.1 Problem 2 a

Find the probability that a randomly chosen person: has a speeding ticket given they do not have a red car.

$S_t \rightarrow$ event: has a speeding ticket

$R \rightarrow$ event: has a red car

$$\mathbb{P}\left(S_t \mid R^c\right) = \frac{\mathbb{P}\left(S_t \cap R^c\right)}{\mathbb{P}\left(R^c\right)} = \frac{45}{515} = \frac{9}{103} \approx 8,738\%$$

The probability that a person has a speeding ticket given they do not have a red car is approximately $8,738\%$.

14.2 Problem 2 b

Find the probability that a randomly chosen person: has a red car given they have a speeding ticket.

$$\mathbb{P}\left(R \mid S_t\right) = \frac{\mathbb{P}\left(R \cap S_t\right)}{\mathbb{P}\left(S_t\right)} = \frac{15}{60} = \frac{1}{4} = 25\%.$$

The probability that a person has a red car given they have a speeding ticket is 25%.

15 Question 1

15.1 Introduction

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. Essentially, it seeks to find a linear function (in the two dimensional case, a line) that best represents the observed data, minimizing the distance between the actual points and those predicted by the line. Its simplicity and explainability make it one of the most commonly used tools in data analysis, applied in fields such as economics, social sciences, and machine learning to make predictions, perform classification, and understand the influence of independent variables on the final outcome. In the following case, the authors of the paper "Linear Regression for Face Recognition"[8] used the technique of linear regression to solve the problem of face recognition.

15.2 Problem Description

Facial recognition is a complex challenge in the field of artificial intelligence and machine learning, as it requires the analysis of high dimensional data to identify or verify a human face among many. This process involves not only extracting unique visual features from an image, such as face shape and the distance between the eyes, nose, and mouth, but also comparing these features with a potentially vast database. From a computational standpoint, facial recognition can be extremely intensive. Each image must be preprocessed, often resized and normalized, in order to extract the relevant features, a process that can require the use of complex mathematical models. Additionally, as the size of reference datasets grows, the need for computational resources increases exponentially. Traditional classification approaches, such as neural networks, can become too time consuming and resource intensive. In this context, linear regression represents an efficient solution, as it simplifies the problem by reducing it to a linear relationship between the extracted facial features and the identities of faces in the database. Although this approach may seem less sophisticated than techniques like convolutional neural networks, it has the advantage of being less computationally expensive while still yielding good results in certain applications. This makes it particularly useful in environments with limited computational resources or where faster processing times are required.

15.3 Proposed Solution

In particular, the solution proposed by the authors of the paper is Linear Regression Classification, where linear regression is used to represent a test image as a linear combination of the training images of a specific class (for example, images of a single person). The underlying idea is that images of the same person lie on a linear subspace, so linear regression is used to build a model of this subspace for each class. Furthermore, the authors highlight how linear regression can also be exploited as an alternative to more traditional dimensionality reduction methods, such as Principal Component Analysis. In the case of linear regression, the images are resized and represented as low dimensional vectors before being used in the regression model. This helps reduce computational complexity and avoid the "curse of dimensionality", which is common when working with high resolution images. The model is initially trained with the training images of each class (person), which are transformed into vectors and organized into a matrix representing the class's linear subspace. Then, to test the model, a test image is provided, and the system uses linear regression to compute a linear combination of the class's training images, attempting to reconstruct the test image. The distance between the reconstructed image and the original test image is then calculated. The class that manages to reconstruct the test image with the lowest error (for example the smallest distance) is chosen as the classification result. The model was subsequently tested on five standard databases, namely: AT&T (in the case of this dataset, two different evaluation protocols were also used), Georgia Tech, FERET, Extended Yale B, and AR. The experiments have shown that the approach based on linear regression is effective in handling variations in facial expressions, poses, and lighting conditions. Moreover, it is particularly useful in addressing the problem of occlusions (for example, when parts of the face are covered by glasses or scarves), offering results comparable to or better than other state of the art facial recognition techniques.

15.4 Conclusion

In conclusion, the approach presented by the authors proves to be both simple and effective. It is particularly well suited for handling occlusions and facial variations by focusing on the correct parts of an image and disregarding the damaged ones. The Linear Regression Classification method demonstrates itself to be competitive with cutting edge techniques, offering a robust solution and significantly improving performance in challenging conditions such as occlusions and variations in facial expressions.

16 Question 3

The iodine value (x) is the amount of iodine necessary to saturate a sample of 100 g of oil.

x	132	129	120	113.2	105	92	84	83.2	88.4	59	80	81.5	71	69.2
y	46	48	51	52.1	54	52	59	58.7	61.6	64	61.4	54.6	58.8	58

Find the regression line that will best fit the data.

$\sum_i x_i = 1307,5 \quad \bar{x} = 93,3929$

$\sum_i y_i = 779,2 \quad \bar{y} = 55,6571$

$$\begin{aligned}\sum_i (x_i - \bar{x})^2 = {} & 38,6071^2 + 35,6071^2 + 26,6071^2 + 19,8071^2 + 11,6071^2 \\ & + (-1,3929)^2 + (-9,3929)^2 + (-10,1929)^2 + (-4,9929)^2 \\ & + (-34,3929)^2 + (-13,3929)^2 + (-11,8929)^2 + (-22,3929)^2 \\ & + (-24,1929)^2 = 6906,6645\end{aligned}$$

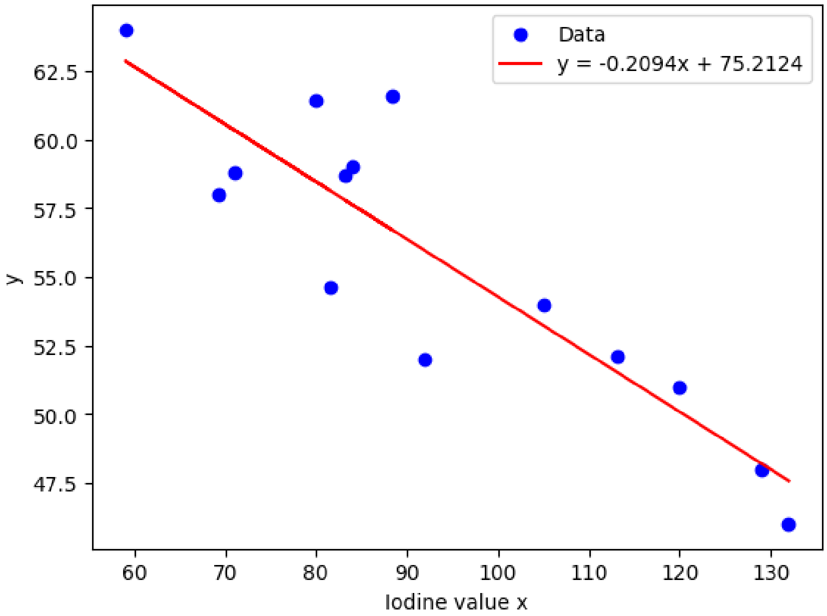
$$\begin{aligned}\sum_i (x_i - \bar{x})(y_i \cdot \bar{y}) = & (38,6071 \cdot (-9,6571)) + (35,6071 \cdot (-7,6571)) + \\ & (26,6071 \cdot (-4,6571)) + (19,8071 \cdot (-3,5571)) + \\ & (11,6071 \cdot (-1,6571)) + (-1,3929 \cdot (-3,6571)) + \\ & (-9,3929 \cdot 3,3429) + (-10,1929 \cdot 3,0429) + \\ & (-4,9929 \cdot 5,9429) + (-34,3929 \cdot 8,3429) + \\ & (-13,3929 \cdot 5,7429) + (-11,8929 \cdot (-1,0571)) + \\ & (-22,3929 \cdot 3,4429) + (-24,1929 \cdot 2,6429) = -1438,3900\end{aligned}$$

$\theta_1 = \frac{-1438,39}{6906,6645} = -0,2083$

$\theta_0 = \bar{y} - \theta_1\bar{x} = 55,6571 + 0,2083 \cdot 93,3929 = 75,0208$

$\hat{y} = \theta_1 x + \theta_0 = -0,2083x + 75,0208$

In the following figure[16], the drawing of the regression line $\hat{y}$ can be seen, the graph was made with the python code attached in the jupyter notebook file relating to the first question.



Figur 3: The regression line

17 Question1

The exponential distribution is given by:

$p(x;\eta) = \frac{1}{\eta}e^{\frac{-x}{\eta}}$

Estimate η using MLE.

Likelihood function creation:

$L(x_1, x_2 \ldots x_n;\eta) = \prod_{i=1}^{n} p\left(x_i;\eta\right) = \prod_{i=1}^{n} \frac{1}{\eta}e^{-\frac{x_i}{\eta}} = \frac{1}{(\eta)^n}e^{-\frac{1}{\eta}\sum_{i=1}^{n} x_i}.$

Compute log $L(x_1, x_2 \ldots x_n;\eta)$ to get Log-Likelihood Function:

$$l(\eta) = \ln L\left(x_1, x_2 \ldots x_n;\eta\right) = \ln\left(\frac{1}{(\eta)^n}e^{-\frac{1}{\eta}\sum_{i=1}^{n} x_i}\right) = -n\ln(\eta) - \frac{1}{\eta}\sum_{i=1}^{n} x_i$$

Maximize Log-Likelihood by computing the derivative and setting it to zero:

$\frac{\delta l}{\delta\eta} = -\frac{n}{\eta} + \frac{1}{\eta^2}\sum_{i=1}^{n} x_i = 0$

$-n\eta + \sum_{i=1}^{n} x_i = \theta$

$\hat{\eta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

18 Question2

Let $X1, X2, X3, ..., Xn$ be a random sample from a Geometric distribution (θ), where θ is unknown. Find the maximum likelihood estimator (MLE) of θ based on this random sample.

Hint: Geometric distribution : $P(x, \theta) = \theta(1-\theta)^{(x-1)}$

Likelihood function creation:

$L(x_1, \dots x_n; \theta) = \prod_{i=1}^{n} P\left(x_i; \theta\right) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i - 1} = \theta^n (1-\theta)^{\sum_{i=1}^{n}(x_i - 1)}$

Compute log $L(x_1, \dots x_n; \theta)$ to get Log-Likelihood Function:

$$\begin{aligned} l\left(x_1, \dots x_n; \theta\right) &= \log\left(L\left(x_1, \dots x_n; \theta\right)\right) = \\ &= n\log(\theta) + \left(\sum_{i=1}^{n}(x_i - 1)\right)\log(1-\theta) \end{aligned}$$

Maximize Log-Likelihood by computing the derivative and setting it to zero:

$\frac{\delta l}{\delta \theta} = \frac{n}{\theta} - \frac{\sum_{i=1}^{n}(x_i - 1)}{1-\theta} = 0$

$\frac{n}{\theta} = \frac{\sum_{i=1}^{n}(x_i - 1)}{1-\theta}$

$n(1-\theta) = \theta\sum_{i=1}^{n}(x_i - 1)$

$n - n\theta = \theta\sum_{i=1}^{n} x_i - n\theta$

$n = \theta\sum_{i=1}^{n} x_i$

$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^{-1} = \frac{1}{\bar{x}}$

19 Question3

Suppose that the lifetime of Polar brand light bulbs is modeled by an exponential distribution $\left(f\left(x_i\right) = \lambda \cdot e^{(-\lambda x_i)}\right)$ with (unknown) parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ?

$X = \{2, 3, 1, 3, 4\}$

Likelihood function creation:

$L(X, \lambda) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i} = \lambda^5 e^{-\lambda \sum_{i=1}^{5} x_i} = \lambda^5 e^{-13\lambda}$

Compute log $L(X, \lambda)$ to get Log-Likelihood Function:

$l(X, \lambda) = \ln(L(X, \lambda)) = 5\ln(\lambda) - 13\lambda$

Maximize Log-Likelihood by computing the derivative and setting it to zero:

$\frac{\delta l(\lambda)}{\delta \lambda} = \frac{5}{\lambda} - 13 = 0 \quad \hat{\lambda} = \frac{5}{13} \approx 0,384615$

Here are the handwritten exercises attached in the following pages, in case anything was unclear.

20 Question 1

20.1 Problem 1 a

The teacher thinks the average height of his students has increased. The average height of a 8th grader ten years ago was 145 cm with a standard deviation of 18 cm. She takes a random sample of 200 students and finds that the average height of his sample is 149 cm. Are students now taller than they were before? Conduct a single-tailed hypothesis test using a 0.05 significance level to evaluate the null and alternative hypotheses.

$N = 200$
$\bar{x} = 149$ cm
$\sigma = 18$ cm
$a = 0.05$
$H_0 : \mu = \mu_0 = 145$ cm
$H_1 : \mu > \mu_0 = 145$ cm

$Z^* = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{N}}} = \frac{149-145}{\frac{18}{\sqrt{200}}} = \frac{20\sqrt{2}}{9} \approx 3,1427$

$Z_{0.05,Right} = 1,645$

$Z^* > Z_{0.05,Right}$

We reject the null hypothesis H_0, there is statistical evidence to state that the average height of students has increased compared to 10 years ago.

20.2 Problem 1 b

He now takes new group of 100 students whose average height is 147 cm. Conduct a single-tailed hypothesis test on this group using a 0.05 significance level to evaluate the null and alternative hypotheses.

$N = 100$
$\bar{x} = 147$ cm
$\sigma = 18$ cm
$a = 0.05$
$H_0 : \mu = \mu_0 = 145$ cm
$H_1 : \mu > \mu_0 = 145$ cm

$Z^* = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{N}}} = \frac{147-145}{\frac{18}{\sqrt{100}}} = \frac{2}{1,8} \approx 1,1111$

$Z_{0.05,Right} = 1,645$

$Z^* < Z_{0.05,Right}$

We can not reject the null hypothesis H_0, there is not enough statistical evidence to state that the average height of students has increased in this case.

21 Question 2

A company manufactures batteries and claims they will last an average of 350 hours under normal use. 30 batteries from the production line are randomly selected and tested. The tested batteries had a mean life span of 325 hours with a standard deviation of 50 hours. Do we have enough evidence to suggest that the claim of an average lifetime of 350 hours is false?

$N = 30$
$\mu_0 = 350$ h
$\bar{x} = 325$ h
$S = 50$ h
$a = 0.05$ (I assumed the value of a since it is not explicitly stated in the instructions)
$H_0 : \mu = \mu_0 = 350$ h
$H_1 : \mu < \mu_0 = 350$ h

$$t^* = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{N}}} = \frac{325-350}{\frac{50}{\sqrt{30}}} = -\frac{\sqrt{30}}{2} \approx -2,7386$$

$t_{0.05,29,Left} = -1,699$

$t^* < t_{0.05,29,Left}$

We reject the null hypothesis H_0, there is enough evidence to conclude that the average battery life is less than 350 hours. In other words, the test suggests that the actual average battery life is likely lower than what the company claims.

22 Question 4

Regression equation: Annual expenditure $= 0.45 * Salary + 10$				
Predictor	Coef	SE Coef	T	P
Constant	10	2.7	5.0	0.00
Salary	0.45	0.197	2.29	0.01

22.1 Problem 4 a

A survey of 102 people on their expenses and salary gives the following statistics. Is there a significant linear relationship between annual salary and expenses? Use a 0.05 level of significance.

The coefficient $Coef$ represents the expected change in the dependent variable (in this case, annual expenses) in response to a one unit change in the independent variable (in this case, salary), holding all other variables in the model constant.

$Coef = 0.45$
$SECoef = 0.197$
$T = 2.29$
$P = 0.01$
$a = 0.05$
$H_0 : Coef = 0$ (No linear relationship)
$H_1 : Coef \neq 0$ (Significant linear relationship)

The P-value for the Salary is 0.01, so we can directly compare it with the level of significance a, that is 0.05:

$P = 0.01 < 0.05 = a$

Since the P-value is below the 0.05 significance level, we reject the null hypothesis H_0. There is sufficient evidence to indicate a significant linear relationship between annual salary and annual expenditure at the 0.05 significance level.

22.2 Problem 4 b

What level of significance will produce a result opposite to what you obtained in the problem 4 a?

To fail to reject the null hypothesis, which states that there is no linear relationship between salary and annual expenditure, we would need to use a significance level more stringent than the P-value of 0.01. Practically, we have to set the level of significance a lower than the P-value of 0.01, for example $a = 0.005$. In this case, we would fail to reject the null hypothesis H_0. This means that there would not be sufficient evidence to conclude a significant linear relationship between salary and annual expenditure at the $a = 0.005$ level of significance.

23 Question 1

You've made different batches of chocolate cookies with varying amounts of butter and chocolate, and recorded which batches sold-out.

	butter	chocolate	sold-out
Batch 1	0.7	0.8	1
Batch 2	0.2	0.45	0
Batch 3	0.1	0.9	0
Batch 4	0.23	0.17	0
Batch 5	0.9	0.75	1

23.1 Problem 1 a

Your ML model M1 gives the following prediction: [1,0,1,1,1]. What is the Empirical Risk here? Assume a 0-1 loss, that is loss is 1 when prediction is wrong and 0 otherwise.

Solution

By $f_{M_1}(b,c)$ we mean the function that generates the set of predictions of the model M_1, which in this case is the vector:

$f_{M_1}(B,C) = [1,0,1,1,1]$

Y is the vector of the real Sold-Out values given in the initial table

$n = 5$

We then calculate the empirical risk:

$$\begin{aligned}\mathcal{L}_{M_1} &= \frac{1}{n}\sum_{i=1}^{5} l\left(y_i, f_{M_1}\left(b_i, c_i\right)\right) = \\ &= \frac{1}{5}(0+0+1+1+0) = \frac{2}{5} = 0.4\end{aligned}$$

23.2 Problem 1 b

Another model M2 is of the form $f(butter, chocolate) = 1$ if ratio of butter to chocolate is more than half, i.e. $\frac{b}{c} > 0.5$. What is the Empirical Risk in that scenario? Use the same loss as above.

Solution

We need to calculate all the results of the function $f(butter, chocolate)$, from now on called $f_{M_2}(b,c)$, obtaining the vector of predictions that we find in the Sold-Out column of the following table.

$f_{M_2}(B,C) = [1,0,0,1,1]$

f_{M_2}	Sold-Out
0,875	1
4/9	0
1/9	0
23/17	1
1,2	1

Now we can actually calculate the empirical risk of the M_2 model.

$$\begin{aligned}\mathcal{L}_{M_2} &= \frac{1}{n}\sum_{i=1}^{5} l\left(y_i, f_{M_2}\left(b_i, c_i\right)\right) = \\ &= \frac{1}{5}(0+0+0+1+0) = \frac{1}{5} = 0.2\end{aligned}$$

23.3 Problem 1 c

Propose your own model (different from M1 and M2), and also propose a loss function such that "not selling out"has heavier penalty and calculate the Empirical Risk.

Solution

We can propose different models that help to minimize the loss function, in this case we show two:

$$M_3 = \begin{cases} 1 & \text{if } 0.5 < \frac{b}{c} < 1.25 \\ 0 & \text{otherwise} \end{cases}$$

$$M_4 = \begin{cases} 1 & \text{if } b + c > 1 \\ 0 & \text{otherwise} \end{cases}$$

We will therefore have that the prediction vectors of the two models are respectively:

$f_{M_3}(B, C) = [1, 0, 0, 0, 1]$

$f_{M_4}(B, C) = [1, 0, 0, 0, 1]$

The empirical risk of the two new models is therefore:

$$\begin{aligned}\mathcal{L}_{M_3} &= \frac{1}{n}\sum_{i=1}^{5} l\left(y_i, f_{M_3}\left(b_i, c_i\right)\right) = \\ &= \frac{1}{5}(0+0+0+0+0) = 0\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{M_4} &= \frac{1}{n}\sum_{i=1}^{5} l\left(y_i, f_{M_4}\left(b_i, c_i\right)\right) = \\ &= \frac{1}{5}(0+0+0+0+0) = 0\end{aligned}$$

The previous loss function was simply defined as a 0-1 loss, that is loss is 1 when prediction is wrong and 0 otherwise. The following table specifies all cases.

Loss Function 1: l_1	True Sold-Out	True Not Sold-Out
Predicted Sold-Out	0	1
Predicted Not Sold-Out	1	0

Let's now define a new loss function that punishes overproduction more and potential lost sales less. Therefore we want to avoid the model preaching the production of a product that will not be sold. This function therefore has three cases instead of two:

1. the model predicts the correct value, in this case $l_2 = 0$
2. the model predicts that the product will not be sold out but in reality the product is completely sold out, in this case $l_2 = 1$
3. the model predicts that the product will be sold out but in reality the product is not sold, in this case $l_2 = 2$

The following table specifies all cases for the Loss Function l_2.

Loss Function 2: l_2	True Sold-Out	True Not Sold-Out
Predicted Sold-Out	0	2
Predicted Not Sold-Out	1	0

Models M_3 and M_4 do not experience any changes in the value of empirical risk since all predictions are correct, in fact both models still have empirical risk equal to zero. By recalculating the empirical risk of the M_1 and M_2 models, we can observe some changes due to the new loss function that penalizes overproduction more.

$$\begin{aligned}\mathcal{L}_{M_1} &= \frac{1}{n}\sum_{i=1}^{5} l_2\left(y_i, f_{M_1}\left(b_i, c_i\right)\right) = \\ &= \frac{1}{5}(0+0+2+2+0) = \frac{4}{5} = 0.8\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{M_2} &= \frac{1}{n}\sum_{i=1}^{5} l_2\left(y_i, f_{M_2}\left(b_i, c_i\right)\right) = \\ &= \frac{1}{5}(0+0+0+2+0) = \frac{2}{5} = 0.4\end{aligned}$$

23.4 Problem 1 d

If the true distribution looks like:

$p(\frac{b}{c} < 0.3) = 0.25,$

$p(0.3 \leq \frac{b}{c} \leq 0.5) = 0.2,$

$p(0.5 < \frac{b}{c} \leq 0.75) = 0.35,$

$p(\frac{b}{c} > 0.75) = 0.2$

The true labels come from a function: $h = 1$ if $\frac{b}{c} > 0.8, 0$ otherwise.

What is the true risk with:

1. 0-1 loss and M2,
2. your proposed loss function and model.

Solution

True risk is calculated as follows:

$$R_{True}(f) = \iint P(x, y) \cdot l(f(x), y) dxdy$$

We need to analyze the true label and prediction values of the M_2 model in different probability bands.

$p(\frac{b}{c} < 0.3) = 0.25 \quad h = 0 \quad M_2 = 0$ **Correct Prediction**

$p(0.3 \leq \frac{b}{c} \leq 0.5) = 0.2 \quad h = 0 \quad M_2 = 0$ **Correct Prediction**

$p(0.5 < \frac{b}{c} \leq 0.75) = 0.35 \quad h = 0 \quad M_2 = 1$ **Wrong Prediction**

$p(\frac{b}{c} > 0.75) = 0.2$ This last band should be divided into two smaller bands.

Dividing the last band I assume that the probability distribution is $\frac{1}{4}$ of the original probability for the section between 0.75 and 0.8, while $\frac{3}{4}$ for the section greater than 0.8. If the distribution within the range were not uniform or if the probability were assigned differently, the value of the true risk could change.

$p(0.75 < \frac{b}{c} \leq 0.8) = \frac{1}{4} \cdot 0.2 = 0.05 \quad h = 0 \quad M_2 = 1$ **Wrong Prediction**

$p(\frac{b}{c} > 0.8) = \frac{3}{4} \cdot 0.2 = 0.15 \quad h = 1 \quad M_2 = 1$ **Correct Prediction**

The true risk for the M_2 model is:

$$R_{True}(f_{\mathrm{M}_2}) = \sum_i p(b, c) \cdot l_1(f_{\mathrm{M}_2}(b_i, c_i), h(b_i, c_i)) =$$
$$= 0 + 0 + 0.35 \cdot 1 + 0.05 \cdot 1 + 0 = 0.40$$

In the second case, we must evaluate the M_3 model using the second loss function l_2 hypothesized in the previous exercise. Again we need to analyze the true label and prediction values of the M_3 model in different probability bands.

$p(\frac{b}{c} < 0.3) = 0.25 \quad h = 0 \quad M_3 = 0$ **Correct Prediction**

$p(0.3 \leq \frac{b}{c} \leq 0.5) = 0.2 \quad h = 0 \quad M_3 = 0$ **Correct Prediction**

$p(0.5 < \frac{b}{c} \leq 0.75) = 0.35 \quad h = 0 \quad M_3 = 1$ **Wrong Prediction**

$p(\frac{b}{c} > 0.75) = 0.2$ This last band should be divided into three smaller bands.

Dividing the last band I assume that the probability distribution is $\frac{1}{4}$ of the original probability for the section between 0.75 and 0.8, while $\frac{2}{4}$ for the section between 0.8 and 1.25, and $\frac{1}{4}$ for the section greater than 1.25. If the distribution within the range were not uniform or if the probability were assigned differently, the value of the true risk could change.

$p(0.75 < \frac{b}{c} \leq 0.8) = \frac{1}{4} \cdot 0.2 = 0.05 \quad h = 0 \quad M_3 = 1$ **Wrong Prediction**

$p(0.8 < \frac{b}{c} < 1.25) = \frac{2}{4} \cdot 0.2 = 0.1 \quad h = 1 \quad M_3 = 1$ **Correct Prediction**

$p(\frac{b}{c} \geq 1.25) = \frac{1}{4} \cdot 0.2 = 0.05 \quad h = 1 \quad M_3 = 0$ **Wrong Prediction**

The true risk for the M_3 model is:

$$\begin{aligned} R_{True}\left(f_{\mathrm{M}_3}\right) &= \sum_i p(b,c) \cdot l_2\left(f_{\mathrm{M}_3}(b_i,c_i), h(b_i,c_i)\right) = \\ &= 0 + 0 + 0.35 \cdot 2 + 0.05 \cdot 2 + 0 + 0.05 \cdot 1 = 0.85 \end{aligned}$$

24 Problem 1

Assume we have N samples, $x_1, ..., x_N$ independently drawn from a normal distribution with known variance σ^2 and unknown mean μ.

24.1 Problem 1 a

Derive the MLE estimator for the mean μ.

Solution

$f\left(x_i \mid \mu\right)=\frac{1}{\sqrt{2 \pi \sigma^2}} e^{-\frac{\left(x_i-\mu\right)^2}{2 \sigma^2}}$

Likelihood function creation:

$L\left(x_1, \ldots, x_N ; \mu\right)=\prod_{i=1}^N f\left(x_i \mid \mu\right)=\prod_{i=1}^N \frac{1}{\sqrt{2 \pi \sigma^2}} e^{-\frac{\left(x_i-\mu\right)^2}{2 \sigma^2}}$

Compute log $L(x_1, \ldots x_n; \mu)$ to get Log-Likelihood Function:

$$\begin{aligned} l(\mu)=\ln L\left(x_1, \ldots, x_N ; \mu\right) &=\sum_{i=1}^N \ln \left(f\left(x_i \mid \mu\right)\right)= \\ &=\sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2 \pi \sigma^2}} e^{-\frac{\left(x_i-\mu\right)^2}{2 \sigma^2}}\right)= \\ &=\sum_{i=1}^N\left(\ln \left(2 \pi \sigma^2\right)^{-\frac{1}{2}}-\frac{\left(x_i-\mu\right)^2}{2 \sigma^2}\right)= \\ &=-\frac{N}{2} \ln \left(2 \pi \sigma^2\right)-\frac{1}{2 \sigma^2} \sum_{i=1}^N\left(x_i-\mu\right)^2 \end{aligned}$$

Maximize Log-Likelihood by computing the derivative and setting it to zero:

$\frac{d l(\mu)}{d \mu}=-\frac{1}{\sigma^2} \sum_{i=1}^N\left(x_i-\mu\right)=0$

$\sum_{i=1}^N x_i-N \mu=0$

$\hat{\mu}_{MLE}=\frac{1}{N} \sum_{i=1}^N x_i=\bar{x}$

24.2 Problem 1 b

Now derive the MAP estimator for the mean μ. Assume that the prior distribution for the mean is itself a normal distribution with mean ν and variance β^2

Solution

$\mu \sim \mathcal{N}\left(\nu, \beta^2\right)$

$f(\nu) = \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}$

$L\left(x_1, \ldots x_N; \mu\right) = f\left(x_1, \ldots, x_N \mid \mu\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x_i-\mu\right)^i}{2\sigma^2}}$

$$
\begin{aligned}
\hat{\mu}_{\text{MAP}} &= \arg\max_\mu f\left(\mu \mid x_1, \ldots, x_N\right) = \arg\max_\mu \frac{f\left(x_1, \ldots, x_N \mid \mu\right) f(\mu)}{f\left(x_1, \ldots, x_N\right)} = \\
&= \arg\max_\mu f\left(x_1, \ldots, x_N \mid \mu\right) f(\mu) = \\
&= \arg\max_\mu L\left(x_1, \ldots, x_N; \mu\right) f(\mu)
\end{aligned}
$$

Now we apply the natural logarithm and subsequently calculate the derivative by equating it to zero:

$$
\begin{aligned}
&\arg\max_\mu \ln L\left(x_1, \ldots, x_N; \mu\right) f(\mu) = \\
&= \arg\max_\mu \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}}\right) \cdot \left(\frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}\right) = \\
&= \arg\max_\mu \left(\sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{N} \frac{\left(x_i-\mu\right)^2}{2\sigma^2} + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(\mu-\nu)^2}{2\beta^2}\right)
\end{aligned}
$$

$\frac{d \ln(f(\mu|x_1,\ldots,x_N))}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \left(x_i - \mu\right) - \frac{\mu-\nu}{\beta^2} = 0$

$\frac{\sum_{i=1}^{N} x_i - N\mu}{\sigma^2} - \frac{\mu-\nu}{\beta^2} = 0$

$\frac{\sum_{i=1}^{N} x_i}{\sigma^2} - \frac{N\mu}{\sigma^2} - \frac{\mu}{\beta^2} + \frac{\nu}{\beta^2} = 0$

$\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{\nu}{\beta^2} = +\frac{N\mu}{\sigma^2} + \frac{N}{\beta^2}$

$\mu\left(\frac{N}{\sigma^2} + \frac{1}{\beta^2}\right) = \frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{\nu}{\beta^2}$

$\mu\left(\frac{\beta^2 N + \sigma^2}{\sigma^2\beta^2}\right) = \frac{\beta^2 \sum_{i=1}^{N} x_i + \sigma^2\nu}{\sigma^2\beta^2}$

$\hat{\mu}_{\text{MAP}} = \frac{\beta^2 \sum_{i=1}^{N} x_i + \sigma^2\nu}{\sigma^2\beta^2} \cdot \frac{\sigma^2\beta^2}{\beta^2 N + \sigma^2} = \frac{\beta^2 N\bar{x} + \sigma^2\nu}{\beta^2 N + \sigma^2}$

25 Problem 4

List (with atleast one or two sentence description/reasoning) 4 limitations of the EM algorithm .

Solution

Here are four notable limitations of the Expectation-Maximization (EM) algorithm:

1. **Sensitivity to Initialization:** the EM algorithm can converge to local optima, particularly if the initial parameters (means, variances, and weights) are poorly

chosen. To mitigate this issue, multiple runs or careful initialization techniques, such as K-means or other clustering algorithms, can be employed to find a satisfactory solution. However, this approach can be computationally demanding.

2. **Known Number of Components:** the EM algorithm requires specifying the number of components in advance (such as the number of clusters in a Gaussian Mixture Model). This assumption can often be inaccurate in real world scenarios where the actual number of components is unknown. Testing different values to estimate the correct number can be time consuming and computationally intensive.

3. **Limited Applicability to Model Types:** The EM algorithm assumes that data is generated from a specific statistical model, which may not hold true for all types of distributions or real-world situations. For example, our data may not actually be generated by two Gaussian distributions, yet the EM algorithm relies on this assumption. This limitation makes it less suitable for cases where the underlying data structure diverges significantly from the assumed model.

4. **Susceptibility to Local Optima:** The EM algorithm can become trapped in local optima, converging to suboptimal solutions rather than the global maximum, especially when the likelihood function contains multiple local maxima.

Referanser

[1] M. Hajji, B. Benhala og I. Hamdi, «Survey of Gradient Descent Variants and Evaluation Criteria,» i *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, IEEE, 2024, s. 01–07.

[2] S. Murali, A. Mutapcic, D. Atienza, R. Gupta, S. Boyd og G. De Micheli, «Temperature-aware processor frequency assignment for mpsocs using convex optimization,» i *Proceedings of the 5th IEEE/ACM international conference on Hardware/software codesign and system synthesis*, 2007, s. 111–116.

[3] S. Murali, A. Mutapcic, D. Atienza mfl., «Temperature control of high-performance multi-core platforms using convex optimization,» i *Proceedings of the conference on Design, automation and test in Europe*, 2008, s. 110–115.

[4] M. Nakai og W. Ke, «Review of the methods for handling missing data in longitudinal data analysis,» *International Journal of Mathematical Analysis*, årg. 5, nr. 1, s. 1–13, 2011.

[5] H. Kang, «The prevention and handling of the missing data,» *Korean journal of anesthesiology*, årg. 64, nr. 5, s. 402–406, 2013.

[6] K. K. Vasan og B. Surendiran, «Dimensionality reduction using principal component analysis for network intrusion detection,» *Perspectives in Science*, årg. 8, s. 510–512, 2016.

[7] F. Song, Z. Guo og D. Mei, «Feature selection using principal component analysis,» i *2010 international conference on system science, engineering design and manufacturing informatization*, IEEE, bd. 1, 2010, s. 27–30.

[8] I. Naseem, R. Togneri og M. Bennamoun, «Linear regression for face recognition,» *IEEE transactions on pattern analysis and machine intelligence*, årg. 32, nr. 11, s. 2106–2112, 2010.