Machine Learning & Data Mining (0432028)
WISE 2024-25
Assignment 01: Introduction to ML and Data Mining

Stefania Zourlidou – zourlidou@uni-koblenz.de

October 28, 2024

---

1 Theoretical Problems

1.1 Cost Function

What is the purpose of a cost function in machine learning? How does minimizing the cost function improve the performance of the model?

1.2 Feature Design

Explain the role of feature design in machine learning and provide an example.

1.3 Training Data

Why is it important to have a large and diverse dataset when training a machine learning model? What could happen if the training data is too small or not varied enough?

1.4 Validation Data

Why is it important to use a separate validation set when training a machine learning model? How does validating the model help improve its performance?

2 Practical Problems

2.1 Linear Algebra

Given the matrices

$$A = \begin{pmatrix} 1 & 7 \\ 4 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 5 & 6 \\ 7 & 3 \end{pmatrix}$$

Compute:

1. The sum $C = A + B$
2. The difference $C = A - B$
3. The product $C = AB$
4. The element-wise multiplication of matrices $A \circ B$
5. The transpose of B, denoted as $C = B^T$
6. The determinant of A:
7. The rank of A:
8. The multiplication of B by vector $x = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$:
9. The multiplication of vector $x = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$ by scalar $a = 5$
10. The element-wise product of vector $x = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$ and vector $y = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$
11. The dot product of vectors $x^T y$ where $x = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$ and $y = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$
12. The length of vector $y = \begin{pmatrix} 0 \\ 132 \end{pmatrix}$:
13. The distance between points $x = \begin{pmatrix} 4 \\ 10 \end{pmatrix}$ and $y = \begin{pmatrix} 12 \\ 4 \end{pmatrix}$ (use the L_2 norm):
14. The distance between points $x = \begin{pmatrix} 4 \\ 10 \end{pmatrix}$ and $y = \begin{pmatrix} 12 \\ 4 \end{pmatrix}$ (use the L_∞ norm):

2.2 Statistics

Answer the following questions.

1. Given the dataset 2, 4, 6, 8, 10, calculate the:
   - mean:
   - median:
   - variance:
   - standard deviation:
2. True or False? The mean is always equal to the median in a symmetric distribution.
3. You are given the exam scores of two different classes:
   - Class X: 70, 75, 80, 85, 90
   - Class Y: 60, 70, 80, 90, 100

   Calculate the mean, median, variance, and standard deviation for both classes. Which class has a higher variability in scores?
   - mean X:
   - median X:
   - variance X:
   - standard deviation X:
   - mean Y:
   - median Y:
   - variance Y:
   - standard deviation Y:
   - Class with higher variability in scores:
4. Given the function $f(x) = -x^2 + 4x$, find the **argmax** values of x in the interval $[0, 5]$. What is the maximum of the function in this interval?
5. A company's profit in thousands of dollars is modeled by the function $P(x) = -2x^2 + 12x - 10$, where x is the amount spent on advertising in thousands of dollars. Find the amount spent on advertising that maximizes the company's profit. What is the maximum profit?

2.3 Logarithms

Select whether each logarithmic expression is correct or incorrect.

Correct	**Incorrect**	
		The base of a natural logarithm is 10.
		The logarithm of 1 is always 0, regardless of the base.
		If $b^y = x$, then $\log_b(x) = y$.
		$\log_b(x \cdot y) = \log_b(x) + \log_b(y)$
		$\log_b\left(\frac{x}{y}\right) = \log_b(x) - \log_b(y)$
		$\log_b(x^y) = y \cdot \log_b(x)$
		$\log_b(1) = 1$
		$\log_b(b) = 1$
		$\log_b(b^x) = x$
		$b^{\log_b(x)} = x$
		$\log_{10}(100) = 3$
		$\ln(e^3) = e$
		$\log_b\left(\frac{1}{x}\right) = -\log_b(x)$
		$\log_b(x)$ is undefined for $x = 1$.
		$\log_b(x)$ is undefined for $b = 1$.

3 Programming Problem

Assume you have a sales dataset (download available here) in a CSV file named `sales_data.csv` with the following columns:

- OrderID: Identifier for each order. Each order can have more than one product.
- ProductCode: Product code.
- QuantityOrdered: Number of product units ordered.
- PriceEach: Price per unit.
- OrderDate: Date on which the order was placed.
- City: City where the order was placed.

Complete the following tasks:

(a) Load the sales data into a pandas DataFrame.

(b) Display the first 5 rows of the dataset.

(c) Check for any missing values in the dataset.

(d) Add a new column `TotalSales` that contains the total sales for each product of an order (i.e., QuantityOrdered * PriceEach).

(e) Find the total sales across all orders.

(f) Identify the product that generated the highest total sales (across all orders).

(g) Find the average sales per order.

(h) Determine the city with the highest number of orders.

(i) Plot a bar chart showing total sales per city.

(j) Save the dataframe in a new file with the name `modified_sales_data.csv`.

Also, answer the following questions.

(a) What is the total sales across all orders? (Task 5 (e)):

(b) Which is the product (`ProductCode`) that generated the highest total sales (across all orders)? (Task 6 (f)):

(c) What is the average sales per order? (Task 7 (g)):

(d) Which is the city with the highest number of orders? Give the name without quotation marks or spaces. (Task 8 (h)):

(e) How many unique orders were received from the city in the previous question? (Task 8 (h)):