

# CIVE 7150 - Data-Driven Decision Support for Civil and Environmental Engineering

## Homework 1

Due date: January 31, 2024

### Instructions

- Submit a PDF writeup named “LASTNAME\_HW1.pdf” in Canvas. Please answer all the questions in the PDF. If a problem asks for graphs or other results, please extract the results from the code and include them in the report. The PDF report should have complete answers to all the questions.
- The problem set contains two types of problem as noted at the beginning of the question or sub-question: (1) analytical, denoted by  $[A]$ , and (2) computational, denoted by  $[C]$ . The response to analytical problems should show all works and derivations. The response to computational problems should contain expected results, and the code should be made available.
- Please use Jupyter notebooks in Python and include comments in the code if needed.
- Include a link to your code in the PDF file.

## [A] Problem 1 [Probability and Expectation] 22 points

In a bag there are 50 balls, each having one of three colors: red (25 balls), blue (15 balls), and green (10 balls).

When a ball is drawn from the bag, you get different points for every color:

- Red balls are worth 3 points each.
  - Green balls are worth 5 points each.
  - Blue balls are worth 10 points each.
- (a) Suppose that you draw a ball at random from the bag. What is the expected total point value of the ball you are drawing?
- (b) If  $X$  is the random variable that denotes the number of points you get by drawing a ball at random, what is the variance of  $X$ ?
- (c) Three balls were picked at random from the bag and it turned out to be two red balls and one green ball. Now, you are drawing another ball from the remaining ones. What is the expected total point value of the ball you are drawing?
- (d) Suppose that  $X$  is a random variable, with expectation  $E(X) = 10$  and variance  $Var(X) = 2$ . What is the value of  $E[X(X - 1)]$  ?
- (e) Is it always true that if  $A$  and  $B$  are two independent event, then the events  $\neg A$  and  $\neg B$  are independent? Show that this is true, or provide a counterexample.

## [A] Problem 2 [Conditional Probability and Bayes Theorem] 18 points

*Probability theory and the Bayes theorem serve as the basis of important machine learning techniques such as linear discriminant analysis (LDA), density estimation, and Naive Bayes (NB) classifiers, which are very effective in text classification settings.*

A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase “free money” is used, whereas this phrase is only used in 1% of non-spam emails.

- (a) What is the probability that an email has the phrase “free money”?
- (b) A new email has just arrived, which does mention “free money”. What is the probability that it is spam?
- (c) Another email does not have the phrase “free money”. What is the probability that it is not spam?

### [A] Problem 3 [Matrices and vectors] 20 points

*Since data is represented in matrices, many of the techniques that we will see, from linear regression to dimensionality reduction and neural networks, will boil down to matrix operations. Optimized matrix operations from modern packages allow us to train and test algorithms much faster than in an iterative approach.*

Define the following two matrices of size  $4 \times 4$ :

$$D_1 = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 4 & -2 \\ -1 & 0 & -3 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix} \quad D_2 = \begin{bmatrix} 1 & -1 & 0 & 2 \\ -1 & 0 & 1 & -1 \\ 0 & -2 & 0 & 1 \\ 3 & 0 & -1 & 0 \end{bmatrix}$$

- (a) Are the 4 columns in each of the matrix linearly independent or linearly dependent?
- (b) What is the rank of each matrix? To answer this, for each of the matrices above, find the maximum number of columns that are linearly independent.
- (c) Assume that we define a column vector

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

Which of the products  $\theta D_1$ ,  $D_1 \theta$ ,  $\theta^T D_1$ , and  $D_1 \theta^T$  can be computed? For the ones that can be computed, please write the result and its dimension.

Please perform these computations by hand and show your work. Do not use a calculator for this problem.

### Problem 4 [Matrix transpose and inverse] 20 points

- [A] (a) Anton speaks French and German; Geraldine speaks English, French and Italian; James speaks English, Italian, and Spanish; Lauren speaks all the languages the others speak except French; and no one speaks any other language. Make a matrix  $A = \{a_{ij}\}$  with rows representing the four people mentioned and columns representing the languages they speak. Put  $a_{ij} = 1$  if person  $i$  speaks language  $j$  and  $a_{ij} = 0$  otherwise.

Compute the matrices  $AA^T$  and  $A^T A$ . What does each entry in these two matrices represent?

- [A] (b) A matrix  $A$  is symmetric if  $A^T = A$ . Prove that the inverse of a symmetric matrix is also a symmetric matrix.
- [C] (c) Generate at random 3 matrices of size  $3 \times 3$  and fill each entry with a random integer chosen from -10 to 10. Use an existing package to compute the inverse of each matrix if it exists.

Include in your report:

- Each of the 3 matrices
- The inverses computed with the package
- Compute the product of each matrix with its inverse to check that you obtain the identity matrix.

## **[C] Problem 5 [Average, variance, and correlation] 20 points**

**Dataset:** The dataset for this assignment has been provided. The prediction task is to predict the price of a house (column **price**) given the other features. Please ignore the columns **id** and **date**, as well as the categorical column **zipcode**. File “kc\_house\_data.csv” includes all the records in the dataset. You should use the entire dataset for the assignment. You can also find a Word document including the feature description in the same folder.

In this problem, we will perform some exploratory data analysis using the house price dataset.

- (a) For each feature, write code to compute the average value, the min and max values, as well as its variance.

Which features have the lowest and the highest average? Include the feature name and their average values for the features with the lowest and highest average.

Which features have the lowest and the highest variance? Include the feature name and their variance values for the features with the lowest and highest variance.

- (b) Compute the correlation coefficient of each feature with the response. Include a table with the correlation coefficient of each feature with the response. Which features are positively correlated (i.e., have positive correlation coefficient) with the response? Which feature has the highest positive correlation with the response?
- (c) Were you able to find any features with a negative correlation coefficient with the response?