CPSC 4300/6300: Assignment 2

Instructions

- Please submit your solution by 11:59PM October 3.
- Submissions should be made through **Canvas**.
- Please complete homework individually.
- Please submit a **single PDF/Word file** including your solutions and the code of your solutions. Feel free to use any third-party libraries.
- No AI generation tool is allowed for this assignment.

1. **Linear Regression (30 points, 15 points per question):**
   Note: This question is based on the heapo_cleaned_dataset which can be found in Assignment 2 on Canvas.

   Dataset Background: The dataset used in this question comes from the HEAPO (Heat Pump and Electricity Advanced Profiling Observatory) project, a large-scale energy monitoring initiative conducted in Switzerland. It contains measurements collected between 2018 and 2024, combining household smart meter data with local weather observations.

   The purpose of the HEAPO project is to better understand how weather conditions and building characteristics influence electricity consumption, especially in homes equipped with heat pumps. For this question you will work with a subset of the HEAPO dataset that includes daily electricity consumption (in kilowatt-hours) from a single household, matched with daily average weather data such as temperature, relative humidity, and sunshine duration from a nearby station.

   (a) Train a predictor to predict electricity consumption as follows:

   $$Consumption\ (kWh) = \theta_0 + \theta_1 \times [Temperature\_avg] + \theta_2 \times [Humidity\_avg] + \theta_3 \times [Sunshine\_Hours]$$

   Report the values of $\theta_0, \theta_1, \theta_2, \theta_3$. Briefly describe your interpretation of these values, i.e., what do they represent? Explain these in terms of the features and labels.

   (b) Split the data into two fractions: the first 90% for training, and the remaining 10% testing. Train the model using the training set. What is the model's MSE and R^2 on the training and on the test set? Did it perform too well on the training set than the test set? If yes, what could be the reason?

2. **Logistic Regression (30 points, 15 points per question):**
   Note: This question is based on the Pima Indians Diabetes dataset which can be found in Assignment 2 on Canvas.

   Dataset Background: The dataset used in this assignment comes from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Its purpose is to support research into the prediction of diabetes based on common medical measurements. See Diabetes_description.txt in Assignment 2 on Canvas for more details.

   (a) Train a set of logistic regression models with different values of C (c = 0.01, 0.1, 1, 10, 100) for both L1 and L2 penalties (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html).
   (b) How does increasing or decreasing C affect the model's performance and feature weights?

CPSC 4300/6300: Assignment 2

3. **Model Evaluation (40 points, 10 points per question)**

You are given the following two classification models, M_1 and M_2, trained on the same binary classification problem. Their confusion matrices are provided below.

M1		Predicted Class	
		Class = 1	Class = 0
Actual Class	Class = 1	150	40
	Class = 0	60	250

M2		Predicted Class	
		Class = 1	Class = 0
Actual Class	Class = 1	250	45
	Class = 0	5	200

Cost Matrix:

Actual \ Predicted	Positive	Negative
Positive	-1	100
Negative	1	0

(a) Compute Accuracy of both models.

(b) Compute the total cost for M_1 and M_2 using the cost matrix.

(c) Explain why M_2 achieves higher accuracy but has a worse cost compared to M_1

(d) Give two real-world scenarios where accuracy is a misleading metric. For each, explain which metric(s) you would prefer (e.g., precision, recall, F1-score, cost-sensitive evaluation) and why.