Department of Civil Engineering
The University of Hong Kong

CIVL7018 Data science for civil engineering Spring 2025

Assignment 1

Due: **March 16th, Sunday, 11:59PM HKT**

Instructor: Dr. Jintao Ke
TAs: Taijie Chen *and* Jiguang Wang

- If the problem specifies that you should use Jupyter Notebook, you are expected to print the Jupyter Notebook as PDF with all codes and outputs included, and submit this printed PDF together with your written (or typed) answers.
- In this assignment, problem 6.3 (c) is optional. We do not take the optional problem into the grading process.
- This assignment has 100 points in total, and it contributes 10% to your final grade of this course.

Problem 1 (Single Variable Linear Regression, *10pt*)

You are tasked with analyzing the relationship between air pollutant concentration and the number of respiratory illness cases reported in a metropolitan area. This analysis will help local health authorities understand the potential impact of air quality on public health and make informed decisions about pollution control measures.

You are provided with the following dataset (Table 1) collected over 6 days in a city:

Table 1: Illness-Pollution Dataset

Day	Air Pollutant Concentration ($\mu g/m^3$)	Number of Respiratory Illness Cases
1	20	50
2	35	70
3	45	90
4	50	110
5	60	125
6	70	135

Tasks:

1.Plot the data with Air Pollutant Concentration on the horizontal axis and Number of Respiratory Illness Cases on the vertical axis. Briefly describe any observable trends or patterns (*3pt*). (Jupyter Notebook)

2. Develop a simple linear regression model to predict the number of respiratory illness cases based on air pollutant concentration. Use the least squares method to estimate the

coefficients (slope and intercept) of the regression line. Write down the equation of the regression line (*5pt*). (Hand-written/type AND Jupyter Notebook)

3. If the air pollutant concentration is measured at 65 µg/m^3 on a particular day, predict the number of respiratory illness cases using your regression model (*2pt*). (Hand-written/type)

Problem 2 (Multiple Linear Regression, *10pt*)

Figure 1: Map of BART (from https://www.bart.gov/system-map)

Consider the ridership of the subway system in the San Francisco Bay area, named the Bay Area Rapid Transit (BART). In this problem, we are investigating the relationship between BART ridership[1] and four socio-economic factors[2], which are **the total population near each station, number of households that own 0 vehicles, total employment, and total road network density**. To simplify the problem, we only consider the total inflow for each station. The data is listed in Table 2 below.

Answer the following questions:

1. Build a multi-variable linear regression model and estimate its coefficients using the matrix-form equation (take the four socio-economic variables as input and total inflow as output) (*5pt*). (Hand-written/type)

2. Use Jupyter Notebook and function `LinearRegression` of *scikit-learn* (https://scikit-learn.org/stable/index.html), validate the multivariate linear regression model you obtained in question 1. Please paste your code (screenshot) and the output in your answer (*5pt*). (Jupyter Notebook)

[1] https://www.bart.gov/about/reports/ridership
[2] https://www.epa.gov/sites/default/files/2021-06/documents/epa_sld_3.0_technicaldocumentationuserguide_may2021.pdf

Table 2: Socio-economics data for four selected BART stations

Station	TotPop	AutoOwn0	TotEmp	TotRdDens	TotInflow
Downtown Berkeley	50383	4784	28318	28.7	3459623
12th Street	11084	1664	33120	42.23	3914019
Powell Street	51122	16059	61815	36.3	8100630
Embarcadero	25970	5383	181995	40.15	13460142
MacArthur	29222	2891	23981	31.30	2535732

Problem 3 (Polynomial Regression, *15pt*)

In urban planning and energy management, it is crucial to understand how temperature affects energy consumption. While a linear model may sometimes provide a rough approximation, real-world data often exhibit non-linear relationships. As temperature increases, energy consumption may rise due to increased use of air conditioning but may also decrease beyond a certain point. This suggests a non-linear relationship, which can be better modeled using polynomial regression.

You are provided with energy consumption data collected from a metropolitan area during the summer season. Your task is to analyze the relationship between daily temperature and energy consumption, apply polynomial regression, and make predictions.

The dataset (Table 3) consists of the following observations:

Table 3: Energy consumption - Temperature Dataset

Temperature (°C)	Energy Consumption (MWh)
15	320
18	340
21	410
25	500
28	620
32	700
35	750
38	770
40	740
42	700

Task:

1. Plot the data points with temperature on the x-axis and energy consumption on the y-axis. Briefly describe any observed trends (*2pt*). (Jupyter Notebook)

2. Fit a 2nd-degree polynomial regression model to the data (i.e., a quadratic model). and estimate the coefficients of the polynomial regression equation and write down the model equation (*6pt*). (Hand-written/type AND Jupyter Notebook)

3. If the temperature is expected to reach 30°C, predict the expected energy consumption using your model. and what is the estimated temperature when the energy consumption is predicted to be 750 MWh (*4pt*)? (Hand-written/type)

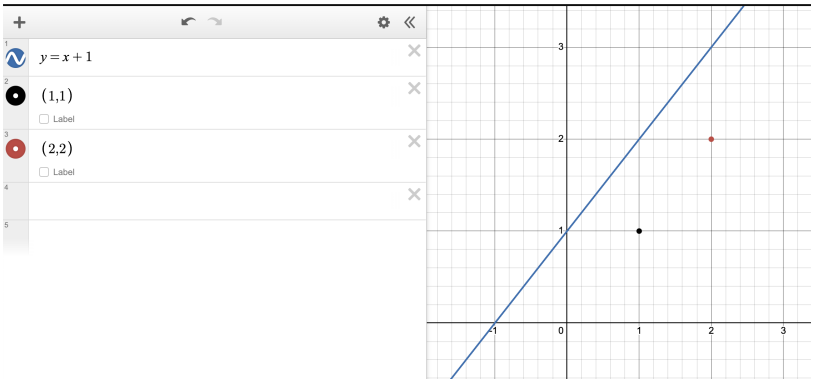
4. Explain the meaning of the coefficients obtained in the regression model (*3pt*). (Hand-written/type)

Problem 4 (Regularization of Regression, *15pt*) We have the following data:

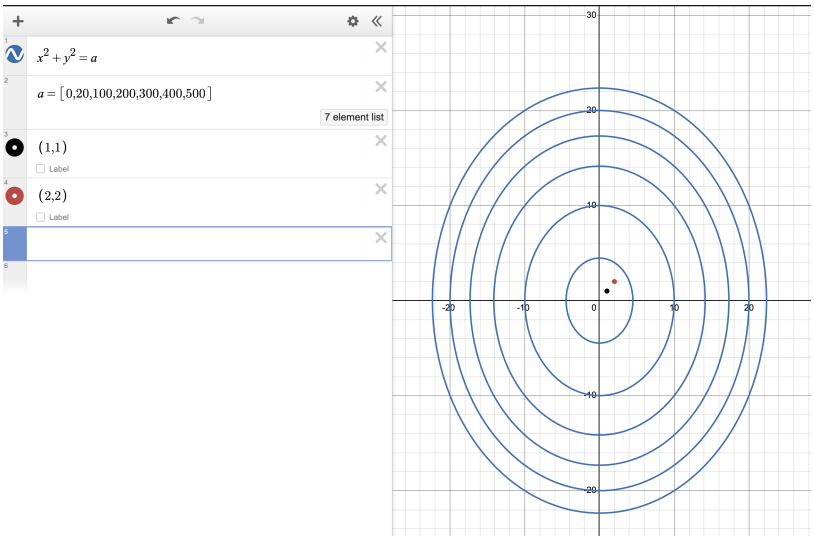
x	y
0.1	1
0.2	2
0.3	3
1	20

Answer the following questions: (Hand-written/type for all subproblems)

1. Consider a linear regression in the form $y = w_0 + w_1 x$, plug the data into the error function of linear regression, and show the simplified expression (*3pt*).
2. Calculate the coefficients w_0 and w_1 (*3pt*).
3. Use the online plotter (https://www.desmos.com/calculator), show the scatter of these four points and the fitted line. Refer to the grammar as follows (*3pt*).



4. Refer to the grammar of the plotter as follows, consider a Ridge regression of $\lambda = 2$, show the contour of (*3pt*)
   (a) the error function of what you obtained in question 1,
   (b) the contour of the regularized error function, and
   (c) the contour of $w_0^2 + w_1^2$.



5. Calculate the coefficients using the regularized error function, and identify the coefficient on the plotter. Can you obtain them without calculation (i.e., graphically)? If yes, how (*3pt*)?

Problem 5 (Logistic Regression, *15pt*) A transportation planning agency is studying the choice of transportation modes (Car, Bus, and Bicycle) by individuals based on specific attributes of the trips. The agency collected the following data on the travel cost (in dollars), travel time (in minutes), and whether the individual chose a particular mode (1 if chosen, 0 otherwise). The dependent variable, "Chosen," indicates whether the individual selected that mode for the trip. The following table provides sample data for the mode "Car": (Hand-written/type for all subproblems)

Table 4: Travel data for the "Car" mode.

Travel Cost ($)	**Travel Time (min)**	**Chosen (1 = Yes, 0 = No)**
5	30	1
10	40	0
8	25	1
15	50	0
3	20	1

1. What are the features of the model, and what is the dependent variable (model output) (*1pt*)?
2. Write the discriminant function for the logistic regression model in terms of the features (travel cost and travel time) and parameters (weights) (*2pt*).
3. Using the discriminant function from Question 1, calculate the probability that an individual chooses a car when the travel cost is \$12 and the travel time is 35 minutes. Assume the initial weights are $w_{\text{cost}} = -0.1$, $w_{\text{time}} = -0.05$, and $b = 0.5$ (*3pt*).

4. Write the formula for the cross-entropy loss for this logistic regression model and calculate the loss for the first row of the table (Travel Cost = 5, Travel Time = 30, Chosen = 1) (*3pt*).

5. Derive the gradient of the cross-entropy loss with respect to the weights (w_{cost}, w_{time}) and bias (*1pt*).

6. Using a learning rate of 0.01, update the weights and bias after one iteration of gradient descent using the first row of the table (*2pt*).

7. Assuming the data in the table represents independent observations, calculate the joint probability of observing the given data under the logistic regression model (*3pt*).

Problem 6 (Support Vector Machine, *15pt*)

1. (Hard Margin SVM, *5pt*) Given a dataset with two positive (with class label $+1$) samples $(1, 4)$ and $(2, 3)$ and three negative (with class label -1) samples $(4, 5)$, $(5, 6)$ and $(5, 5)$. Find the maximum hard margin separating hyperplane of SVM and point out the support vectors. (Hand-written/type AND Jupyter Notebook)

2. (Soft Margin SVM, *6pt*) Consider a soft-margin SVM problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \xi_n \\ \text{s.t.} \quad & y_n(\mathbf{w}^\top \mathbf{x}_n + w_0) \geq 1 - \xi_n, \forall n \in \{1, ..., N\} \end{aligned} \tag{1}$$

   For the three data points marked in Figure 2, what should be the values or value ranges of ξ_n corresponding to them? Are these data points correctly classified? (Hand-written/type)

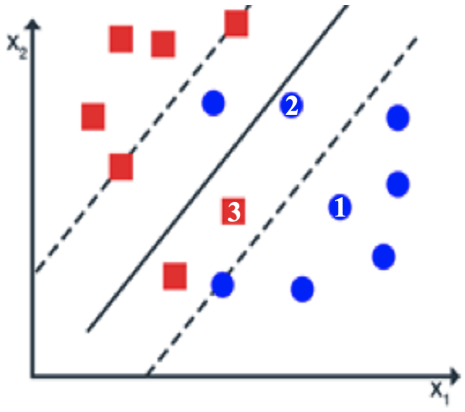


Figure 2: Soft-Margin SVM

3. (Kernel Methods, *4pt*) Generally, we call a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function if there exist a space $\mathcal{F}$ with an inner product $\langle \cdot, \cdot \rangle$ and a feature map $\phi : \mathcal{X} \to \mathcal{F}$

such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. For example $k(\mathbf{x}, \mathbf{y}) = 1 + \mathbf{x}^\top \mathbf{y}$ is a kernel function when we choose $\mathcal{X} = \mathbb{R}^d$, $\mathcal{F} = \mathbb{R}^{d+1}$ and $\phi(\mathbf{x}) = (1, \mathbf{x})$. This kernel lifts the data in $\mathbb{R}^d$ into $\mathbb{R}^{d+1}$. Please use the above definition of kernel functions to show that: (Hand-written/type)

(a) $k(x, y) = (1 + xy)^n$ is a kernel function on $\mathcal{X} = \mathbb{R}$.

(b) $k(x, y) = xy - 1$ is NOT a kernel function on $\mathcal{X} = \mathbb{R}$. (Hint: Use the positive definiteness property of inner products)

(c) (Optional) $k(x, y) = \min(x, y)$ is a kernel function on $\mathcal{X} = [0, 1]$. (Hint: Kernel methods can lift data into an infinite-dimensional space)

Problem 7 (Decision Trees, *5pt*)

Consider a binary classification data set with two features x_1 and x_2. We plan to use a decision tree to classify it. As shown in Figure 3, there are two ways to split it for the first time, using $x_1 = 3.5$ and $x_2 = 3.0$ as the split criteria, respectively. Please compare which split criterion is better. (Hint: Use the Gini index.) (Hand-written/type)

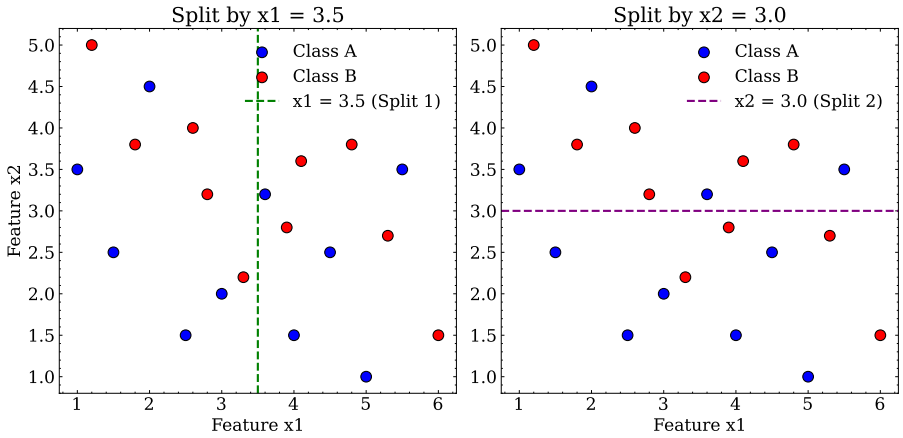


Figure 3: A binary classification dataset and two splits of a decision tree

Problem 8 (Ensemble Learning, *15pt*) A transportation agency is working on predicting whether travelers will use public transportation (Bus) or a private car (Car) for their daily commute. To improve the accuracy of their predictions, they have decided to use an **ensemble learning approach** by combining predictions from three different models:

- **Model 1**: A decision tree classifier trained on travel distance and cost.
- **Model 2**: A logistic regression model trained on travel time and convenience score.
- **Model 3**: A support vector machine (SVM) trained on travel time, cost, and user preferences.

The agency uses a **voting-based ensemble method**, where the final prediction is determined by majority vote among the three models.

The following table shows the predictions from each model for five travelers:

Table 5: Predictions from three models for five travelers.

Traveler	**Model 1 Pred.**	**Model 2 Pred.**	**Model 3 Pred.**	**Ground Truth**
1	Car	Bus	Bus	Bus
2	Bus	Bus	Car	Bus
3	Car	Car	Bus	Car
4	Bus	Bus	Bus	Bus
5	Car	Car	Car	Car

(Hand-written/type for all subproblems)

1. For each traveler, determine the final prediction made by the ensemble model based on majority voting (*3pt*).

2. Calculate the accuracy of the ensemble model based on the predictions and the ground truth (*3pt*).

3. Explain how the ensemble model may affect the bias and variance compared to using a single model (e.g., Model 1) (*3pt*).

4. Suppose the agency decides to use **weighted voting** instead of majority voting, assigning weights to the models as follows:

   - Model 1: 0.2
   - Model 2: 0.5
   - Model 3: 0.3

   For each traveler, determine the final prediction using the weighted voting method (*3pt*).

5. Briefly explain the difference between bagging and boosting in ensemble learning. Which method would be more suitable if the agency wants to reduce overfitting (*3pt*)?