Assignment 1 - Regression and classification

Spring 2024

Due February 29 23:59

- 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 2.2, problem 4, and problem 6.2.4 are optional.** We do not take optional problems into the grading process.
- This assignment has 100 points in total, and it contributes 10% to your final grade of this course.

1 Single Variable Linear Regression (10pt)

Simple linear regression is widely used in the transportation industry, especially by engineering consultants. Imagine that you are an intern engineer and will provide suggestions to the government for the new city arterial project. The government is interested in how traffic density relates to traffic volume (flow). The following data is available to you.

Table 1: Available traffic density and volume data

Traffic Density (veh/km)	Traffic Volume (veh/hour)
10	500
35	1200
18	1000
45	1800

If the predicted traffic using this arterial is 1000 $veh/hour$, what is the estimated traffic density? Draw the data in a graph with density as the horizontal axis and volume as the vertical axis. Build a simple linear regression model and estimate the coefficients using the method of **least squared method**. Present your steps and results. (*Hand-written/type* ***AND*** *Jupyter Notebook*.)

2 Multiple Linear Regression

2.1 A case study (15pt)

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

Consider the ridership of a public transportation system in the San Francisco bay area, the Bay Area Rapid Transit (BART). In this problem, we are investigating the relationship between BART ridership and four socio-economic factors, which are **the total population near each station, number of households that own 0 vehicles, total employment, and total road network density**. The raw census data is collected by the Environment Protection Agency of the United States[1] while the ridership data is available at the BART's official website[2]. For simplicity, we only consider the total inflow for each station. The data is listed in Table 2 below.

1. Using the data given, build a multi-variable linear regression model and estimate its coefficients using Maximum Likelihood Estimation, taking the four socio-economic variables as input and total inflow as output. (*Hand-written/type.*)

2. Using the data given, build a multi-variable linear regression model, taking the four socio-economic variables as input and total inflow as output. (*Jupyter Notebook, using both matrix product and sklearn approach.*)

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

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

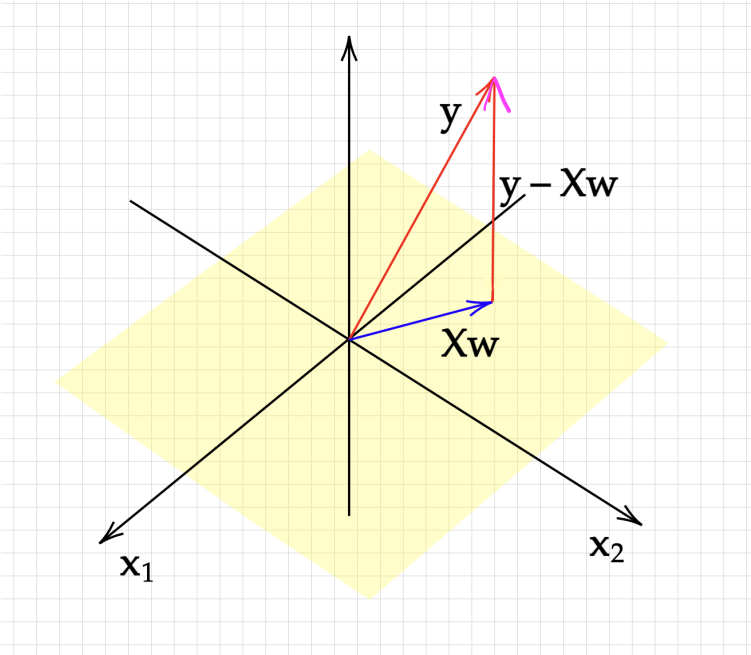
Station	TotPop	AutoOwn0	TotEmp	TotRdDens	TotInflow
Downtown Berkeley	50383	4784	28318	28.7	3459623
12th Street	11084	1664	33120	42.23	3914019
Powell Street	51122	16059	61815	36.3	8100630
Embarcadero	25970	5383	181995	40.15	13460142
MacArthur	29222	2891	23981	31.30	2535732

3. If **the total population near each station, number of households that own 0 vehicles, total employment, and total road network density** for station **Montgomery Street** are 34689, 9443, 148355, 38.9, respectively. What is the estimated total ridership inflow? (*Jupyter Notebook.*)

2.2 Geometric Interpretation of Multiple Linear Regression

This problem is optional.

Consider the observations $\mathbf{X}$ and objective $\mathbf{y}$. We try to fit a multiple linear model with coefficient vector $\mathbf{w}$ that projects $\mathbf{X}$ to $\mathbf{y}$, i.e., $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$. Answer the following questions. (This question requires some basic linear algebra knowledge. *Hand-written/type.*)



1. Answer **True** or **False**. $\mathbf{X}\mathbf{w} \in$ the column space of $\mathbf{X}$.

2. Answer **True** or **False**. To minimize the norm of *distance* between $\hat{\mathbf{y}}$ and $\mathbf{y}$, $\mathbf{X}$ has to be the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$.

3. Given that the *distance* (technically, it is called the residual vector) between $\hat{\mathbf{y}}$ and $\mathbf{y}$ is $\mathbf{y} - \hat{\mathbf{y}}$. Show that $\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. Tip: If $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to span($\mathbf{X}$), $\mathbf{X}^\top (\mathbf{y} - \hat{\mathbf{y}}) = 0$.

3 Polynomial Regression (10pt)

Recall the scenario in Problem 1. In reality, it is not possible to consider every road has a linear relationship between traffic volume and density. When the density gets higher and higher, the traffic condition may become worse, and thus the volume may decrease.

Some new data were collected and are available as shown below. The government is still interested in how traffic density relates to traffic volume (flow).

Table 3: Traffic density and volume data from Problem 1

Traffic Density (veh/km)	Traffic Volume (veh/hour)
10	500
35	1200
18	800
45	1800

Table 4: Additional Traffic density and volume data

Traffic Density (veh/km)	Traffic Volume (veh/hour)
58	1300
65	1000
82	400
74	800

If the predicted traffic using this arterial is 1000 $veh/hour$, what will be the estimated traffic density? What will be the traffic volume if the density goes to 90 veh/km? Draw the data in a graph with density as the horizontal axis and volume as the vertical axis. Build a 2nd-order polynomial regression model and estimate the coefficients. Present your steps and results. (*Jupyter Notebook*.)

4 Regularization of Regression

This problem is optional.

You are welcome to refer to ChatGPT, but don't forget to cite anything borrowed from it.

1. Finish the in-class exercise on slide 55 of Lec 2. (*Jupyter Notebook*.)
2. If we apply regularization (no matter if it is L1 or L2), is it safe to say that we can avoid (at least reduce) the problem of overfitting? Explain. (*Hand-written/type*.)

5 Logistic Regression (15pt)

A transportation engineering consultant company is evaluating a city corridor is being evaluated by a transportation engineering consultant at different times of the day for its congestion level, by its vehicle density, and flow. If the corridor is identified as congested, the "Evaluated as Congested?" attribute is 1, and is 0 if it is not congested. The data is listed below.

Table 5: Traffic density and volume data and whether it is identified as congested

Time of day	Traffic Density (veh/km)	Traffic Volume (veh/hour)	Evaluated as Congested?
Morning	50	2000	1
Noon	20	900	0
Afternoon	40	1500	1
Evening	10	500	0
Midnight	5	300	0

1. What are the features of the model, and what will be the dependent variable (model output)? What will be the discriminant function?

2. Calculate the probability that when the traffic density is 30 veh/km, volume is 1000 veh/hour, the road is evaluated as congested, using the discriminant function you proposed in question 1.

3. Calculate the joint probability that Table 1 is observed.

4. Consider cross-entropy loss, and calculate the derivative of the loss with respect to all variables in your discriminant function.

5. Initializing all the variables in your discriminant function to be 1, and using the learning rate of 0.01. Calculate the estimated variable after 1 iteration.

6 Support Vector Machine

6.1 (20pt)

Consider a soft-margin support vector machine problem.

$$\min_{\mathbf{w}} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_n \xi_n \tag{1}$$

$$\text{s.t.} \quad y_n(\mathbf{w}^\top \mathbf{x}_n + w_0) \geq 1 - \xi_n \forall n \in \{1, \cdots, N\}, \tag{2}$$

where C is the regularization parameter and ξ_n is the slack for each data point n, which allows the misclassification of data points in the event that the data is not linearly separable.

1. Intuitively, where does a data point lie relative to where the margin is when $\xi_n = 0$? Is this data point classified correctly?

2. Intuitively, where does a data point lie relative to where the margin is when $0 < \xi_n \leq 1$? Is this data point classified correctly?

3. Intuitively, where does a data point lie relative to where the margin is when $\xi_n > 1$? Is this data point classified correctly?

4. Match the scenarios as drawn in Figure 1 to the problems below (answer left or right). The dark squares and circles represent the support vector, while the shallow squares and circles represent successfully identified samples.

   - A soft margin SVM with $C = 0.1$.
   - A soft margin SVM with $C = 10$.

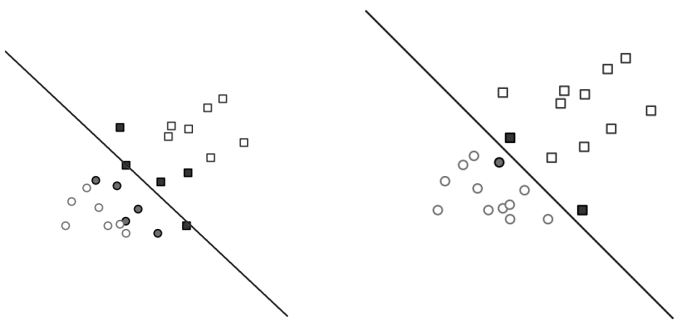


Figure 2: Scenarios

6.2 (15pt)

Consider a hard margin support vector machine problem:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}||\mathbf{w}||^2 \tag{3}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top x_i - b) \geq 1 : \forall i, \tag{4}$$

where $\mathbf{w}$ is the coefficient vector of the hyperplane $\mathcal{H} = \{x : \mathbf{w}^\top \mathbf{x} - b = 0\}$. Imagine that there are multiple points in a space, and they are linearly separatable by the hyperplane.

1. Explain the meaning of a support vector in plain words.
2. Consider two points on $\mathcal{H}$, $\mathbf{x}_0$ and $\mathbf{x}_1$. Show that $(\mathbf{x}_1 - \mathbf{x}_0) \perp \mathbf{w}$.
3. Express the distance of any arbitrary point $\mathbf{z}$ to the hyperplane in the matrix (or vector) product form. (Hint: consider the dot product of vectors.)
4. (*This subproblem is optional*) Derive the support vector machines problem as shown at the beginning of this problem using what you obtained in Problem 6.2.3. Suppose that you have a training set of points $\{x_i\}$.

7 Decision Trees (15pt)

1. Show that any binary tree of depth d has at most $2^{d+1} - 1$ nodes (the depth of a decision tree is the length of the longest path from the root to the leaf).

2. What is the lower bound of the depth of a binary decision tree with m nodes?

3. Consider an unconstrained binary decision tree, meaning that each input observation will be matched with one leaf after training. If N samples are input into the tree, discuss the depth of this tree.

This is the end of this assignment.