AIE1901 Assignment 3

Due date: 11:59 PM, Wednesday, November 19, 2025.

Remark:

1) The Maximum point is 100.
2) It is okay to use LLM (such as ORLM, DeepSeek) to help you generate the answer, but it is optional.
3) Attach the computer code, prompt you to use in your submission.
4) Please download `communities_crime_feature_names.csv` and `communities_crime_features.csv` and `communities_crime_target.csv` and put them into the same folder as you are running the code.
5) You need to run the code for second part using your own laptop instead of Google Colab. Please download `cvxpy` and `Mosek` packages into your own computer environment.

Question 1 (Linear Regression). *We have access to the ”Communities and Crime” dataset, which contains $n = 100$ features and $m = 1994$ samples (variable descriptions provided in Appendix in Page 6). This dataset combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. Our objective is to predict how community characteristics impact violent crime rates using linear regression.*

Specifically, you first upload the feature matrix $X \in \mathbb{R}^{m\times n}$, target vector $y \in \mathbb{R}^{m\times 1}$, and the feature name corresponding to each index using the python code below:

```
import numpy as np
# Load feature matrix
features = np.loadtxt(′communities_crime_features.csv′, delimiter=′,′, skiprows=1)
print(features)
# Load target vector y
target = np.loadtxt(′communities_crime_target.csv′, delimiter=′,′, skiprows=1)
print(target)
# Load feature name
feature_names_data = np.loadtxt(′communities_crime_feature_names.csv′,
                                delimiter=′,′, dtype=str, skiprows=1)
print(feature_names_data)
```

We now want to find an coeffient vector $\beta \in \mathbb{R}^{n\times 1}$ to minimize the fitting error, which leads to the following optimization problem:

$$\min_{\beta\in\mathbb{R}^{n\times 1}} \ \|y - X\beta\|_2^2.$$

Denote by the optimal solution as β^*. Then, our predictive model is

$$\hat{y} = \beta_1^* x_1 + \beta_2^* x_2 + \cdots + \beta_n^* x_n.$$

In the equation above, x_i denotes the i-th feature, $\hat{y}$ denotes our estimation of the crime rate.

Tasks:

- *Use `cvxpy` package to find the optimal solution β^*. Provide the screenshot of its value.*
- *Provide the optimal value of the optimization problem, i.e., $\|y - X\beta^*\|_2^2$.*

(50 points)

Solution to Question 1. □

Question 2 (Variable Selection)**.** *From Question 1, we observed that interpreting a model with 100 coefficients is challenging. To enhance interpretability, we now seek a sparse model with at most* 4 *non-zero coefficients. This leads to the optimization problem:*

$$\min_{\beta \in \mathbb{R}^{n\times 1}, \|\beta\|_0 \le 4} \quad \|y - X\beta\|_2^2,$$

where $\|\beta\|_0$ *counts the number of non-zero entries in* β*. We can reformulate this using binary variables* $q_i \in \{0,1\}$*:*

$$\begin{aligned} \min_{\beta \in \mathbb{R}^{n\times 1}, q \in \{0,1\}^n} \quad & \|y - X\beta\|_2^2 \\ \textit{Subject to} \quad & \sum_{i=1}^n q_i \le 4, \\ & -M \cdot q_i \le \beta_i \le M \cdot q_i, \quad i = 1, \dots, n, \end{aligned}$$

where we specify constant $M = 0.4$*. You may observe solving the formulation above is time-consuming and instable, so we add a small regularization:*

$$\begin{aligned} \min_{\beta \in \mathbb{R}^{n\times 1}, q \in \{0,1\}^n} \quad & \|y - X\beta\|_2^2 + \lambda \cdot \|\beta\|_2^2 \\ \textit{Subject to} \quad & \sum_{i=1}^n q_i \le 4, \\ & -M \cdot q_i \le \beta_i \le M \cdot q_i, \quad i = 1, \dots, n, \end{aligned}$$

where we use $\lambda = 5$ *for the regularization parameter. Let's consider the reformulation of this regularized problem:*

$$\begin{aligned} \min_{\beta \in \mathbb{R}^{n\times 1}, q \in \{0,1\}^n, t \in \mathbb{R}^{n\times 1}} \quad & \|y - X\beta\|_2^2 + \lambda \cdot \sum_{i=1}^n t_i \\ \textit{Subject to} \quad & \beta_i^2 \le t_i, \quad i = 1, \dots, n, \\ & \sum_{i=1}^n q_i \le 4, \\ & -M \cdot q_i \le \beta_i \le M \cdot q_i, \quad i = 1, \dots, n. \end{aligned} \tag{1}$$

Tasks:

- *Use* `cvxpy` *package to find the optimal solution* β^* *of Problem* (1)*. Recall we take parameters* $M = 0.4, \lambda = 5$*.*
- *You will find only 4 entries of* β^* *are non-zero. Identify which 4 features were selected and report their coefficients.*
- *Interpret the meaning of these features in the context of crime prediction.*
- *Provide the optimal value of the optimization problem, i.e.,* $\|y - X\beta^*\|_2^2$*. Compare the optimal value with that of Question 1. What do you observe?*

Important Hints:

- *For this question, please use* `Mosek` *solver with detailed verbose output by using the code* `problem.solve(solver=cp.MOSEK, verbose=True)`*. You can download Mosek license by visiting https://www.mosek.com/products/academic-licenses/.*

  You can download Mosek package by visiting https://www.mosek.com/downloads/

(50 points)

Solution to Question 2. □

APPENDIX: VARIABLE INFORMATION

Attribute Information: (100 predictive variables, 1 target variable)

Variable Name	**Description**
state	US state (by number) - not counted as predictive above, but if considered, should be considered nominal
population	Population for community (numeric - decimal)
householdsize	Mean people per household (numeric - decimal)
racepctblack	Percentage of population that is African American (numeric - decimal)
racePctWhite	Percentage of population that is Caucasian (numeric - decimal)
racePctAsian	Percentage of population that is of Asian heritage (numeric - decimal)
racePctHisp	Percentage of population that is of Hispanic heritage (numeric - decimal)
agePct12t21	Percentage of population that is 12-21 in age (numeric - decimal)
agePct12t29	Percentage of population that is 12-29 in age (numeric - decimal)
agePct16t24	Percentage of population that is 16-24 in age (numeric - decimal)
agePct65up	Percentage of population that is 65 and over in age (numeric - decimal)
numbUrban	Number of people living in areas classified as urban (numeric - decimal)
pctUrban	Percentage of people living in areas classified as urban (numeric - decimal)
medIncome	Median household income (numeric - decimal)
pctWWage	Percentage of households with wage or salary income in 1989 (numeric - decimal)
pctWFarmSelf	Percentage of households with farm or self employment income in 1989 (numeric - decimal)
pctWInvInc	Percentage of households with investment / rent income in 1989 (numeric - decimal)
pctWSocSec	Percentage of households with social security income in 1989 (numeric - decimal)
pctWPubAsst	Percentage of households with public assistance income in 1989 (numeric - decimal)
pctWRetire	Percentage of households with retirement income in 1989 (numeric - decimal)
medFamInc	Median family income (differs from household income for non-family households) (numeric - decimal)
perCapInc	Per capita income (numeric - decimal)
whitePerCap	Per capita income for Caucasians (numeric - decimal)
blackPerCap	Per capita income for African Americans (numeric - decimal)
indianPerCap	Per capita income for Native Americans (numeric - decimal)

Variable Name	Description
AsianPerCap	Per capita income for people with Asian heritage (numeric - decimal)
HispPerCap	Per capita income for people with Hispanic heritage (numeric - decimal)
NumUnderPov	Number of people under the poverty level (numeric - decimal)
PctPopUnderPov	Percentage of people under the poverty level (numeric - decimal)
PctLess9thGrade	Percentage of people 25 and over with less than a 9th grade education (numeric - decimal)
PctNotHSGrad	Percentage of people 25 and over that are not high school graduates (numeric - decimal)
PctBSorMore	Percentage of people 25 and over with a bachelor's degree or higher education (numeric - decimal)
PctUnemployed	Percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)
PctEmploy	Percentage of people 16 and over who are employed (numeric - decimal)
PctEmplManu	Percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
PctEmplProfServ	Percentage of people 16 and over who are employed in professional services (numeric - decimal)
PctOccupManu	Percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
PctOccupMgmtProf	Percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal)
MalePctDivorce	Percentage of males who are divorced (numeric - decimal)
MalePctNevMarr	Percentage of males who have never married (numeric - decimal)
FemalePctDiv	Percentage of females who are divorced (numeric - decimal)
TotalPctDiv	Percentage of population who are divorced (numeric - decimal)
PersPerFam	Mean number of people per family (numeric - decimal)
PctFam2Par	Percentage of families (with kids) that are headed by two parents (numeric - decimal)
PctKids2Par	Percentage of kids in family housing with two parents (numeric - decimal)
PctYoungKids2Par	Percent of kids 4 and under in two parent households (numeric - decimal)
PctTeen2Par	Percent of kids age 12-17 in two parent households (numeric - decimal)
PctWorkMomYoungKids	Percentage of moms of kids 6 and under in labor force (numeric - decimal)
PctWorkMom	Percentage of moms of kids under 18 in labor force (numeric - decimal)
NumIlleg	Number of kids born to never married (numeric - decimal)

Variable Name	Description
PctIlleg	Percentage of kids born to never married (numeric - decimal)
NumImmig	Total number of people known to be foreign born (numeric - decimal)
PctImmigRecent	Percentage of immigrants who immigrated within last 3 years (numeric - decimal)
PctImmigRec5	Percentage of immigrants who immigrated within last 5 years (numeric - decimal)
PctImmigRec8	Percentage of immigrants who immigrated within last 8 years (numeric - decimal)
PctImmigRec10	Percentage of immigrants who immigrated within last 10 years (numeric - decimal)
PctRecentImmig	Percent of population who have immigrated within the last 3 years (numeric - decimal)
PctRecImmig5	Percent of population who have immigrated within the last 5 years (numeric - decimal)
PctRecImmig8	Percent of population who have immigrated within the last 8 years (numeric - decimal)
PctRecImmig10	Percent of population who have immigrated within the last 10 years (numeric - decimal)
PctSpeakEnglOnly	Percent of people who speak only English (numeric - decimal)
PctNotSpeakEnglWell	Percent of people who do not speak English well (numeric - decimal)
PctLargHouseFam	Percent of family households that are large (6 or more) (numeric - decimal)
PctLargHouseOccup	Percent of all occupied households that are large (6 or more people) (numeric - decimal)
PersPerOccupHous	Mean persons per household (numeric - decimal)
PersPerOwnOccHous	Mean persons per owner occupied household (numeric - decimal)
PersPerRentOccHous	Mean persons per rental household (numeric - decimal)
PctPersOwnOccup	Percent of people in owner occupied households (numeric - decimal)
PctPersDenseHous	Percent of persons in dense housing (more than 1 person per room) (numeric - decimal)
PctHousLess3BR	Percent of housing units with less than 3 bedrooms (numeric - decimal)
MedNumBR	Median number of bedrooms (numeric - decimal)
HousVacant	Number of vacant households (numeric - decimal)
PctHousOccup	Percent of housing occupied (numeric - decimal)
PctHousOwnOcc	Percent of households owner occupied (numeric - decimal)

Variable Name	**Description**
PctVacantBoarded	Percent of vacant housing that is boarded up (numeric - decimal)
PctVacMore6Mos	Percent of vacant housing that has been vacant more than 6 months (numeric - decimal)
MedYrHousBuilt	Median year housing units built (numeric - decimal)
PctHousNoPhone	Percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal)
PctWOFullPlumb	Percent of housing without complete plumbing facilities (numeric - decimal)
OwnOccLowQuart	Owner occupied housing - lower quartile value (numeric - decimal)
OwnOccMedVal	Owner occupied housing - median value (numeric - decimal)
OwnOccHiQuart	Owner occupied housing - upper quartile value (numeric - decimal)
RentLowQ	Rental housing - lower quartile rent (numeric - decimal)
RentMedian	Rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)
RentHighQ	Rental housing - upper quartile rent (numeric - decimal)
MedRent	Median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)
MedRentPctHousInc	Median gross rent as a percentage of household income (numeric - decimal)
MedOwnCostPctInc	Median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)
MedOwnCostPctIncNoMtg	Median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal)
NumInShelters	Number of people in homeless shelters (numeric - decimal)
NumStreet	Number of homeless people counted in the street (numeric - decimal)
PctForeignBorn	Percent of people foreign born (numeric - decimal)
PctBornSameState	Percent of people born in the same state as currently living (numeric - decimal)
PctSameHouse85	Percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)
PctSameCity85	Percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)
PctSameState85	Percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)
LandArea	Land area in square miles (numeric - decimal)
PopDens	Population density in persons per square mile (numeric - decimal)
PctUsePubTrans	Percent of people using public transit for commuting (numeric - decimal)

Variable Name	Description
LemasPctOfficDrugUn	Percent of officers assigned to drug units (numeric - decimal)

The variable ”ViolentCrimesPerPop” is our target variable, which represents *Total number of violent crimes per 100K population (numeric - decimal).*