Problem Set 2

Oke

CEE 616: Data Mining and Machine Learning for Engineers *10.02.2025*

Due October 14, 2025 at 11:59PM. Submit via Canvas.

The standard problems are worth a total of **93 points**, with **7 extra credit points** available.

Submission instructions

There are two options for submission:

1. JupyterLab Notebook. Please name your notebook as follows:
   `<lastname>-<firstname>-PS2.ipynb`
2. R/Python/MATLAB script *and* PDF document with supporting responses. Your PDF should have complete responses to all the questions (including all the required plots). Your script should be clearly commented, producing all the results and plots you show in your PDF document. The filenames should be in a similar format as described above.

Whenever datasets are provided, be sure to leave the relative path and filenames as originally given in order to ensure that your scripts will run properly. For instance, here, all the data sets are found in the `data` folder. So, when you call `read.csv()`, use the relative path, e.g. `data/Default.csv`. This way, when you submit your work, you need not include the data. I will have the exact same folder and will be able to run all the scripts as the same path will be referenced.

Problem 1 *Logistic regression (20 pts; 2 pts EC)*

Part 1.1

[4] In a binary logistic regression with a multiple predictors $\boldsymbol{x}^\intercal = (1, x_1, \ldots, x_D)$, the logistic function is given by:

$$p(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^\intercal \boldsymbol{x}}} \tag{1}$$

where $\boldsymbol{w}^\intercal = (b, w_1, \ldots, w_D)$. Using this function, show explicitly that the log-odds or logit function of $p(y = 1|\boldsymbol{x})$ is given by:

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \boldsymbol{w}^\intercal \boldsymbol{x} \tag{2}$$

Part 1.2

The figure below shows the scatter plot of `income` versus `balance` from the `Default` dataset (see Lecture 3a for a description of the dataset).

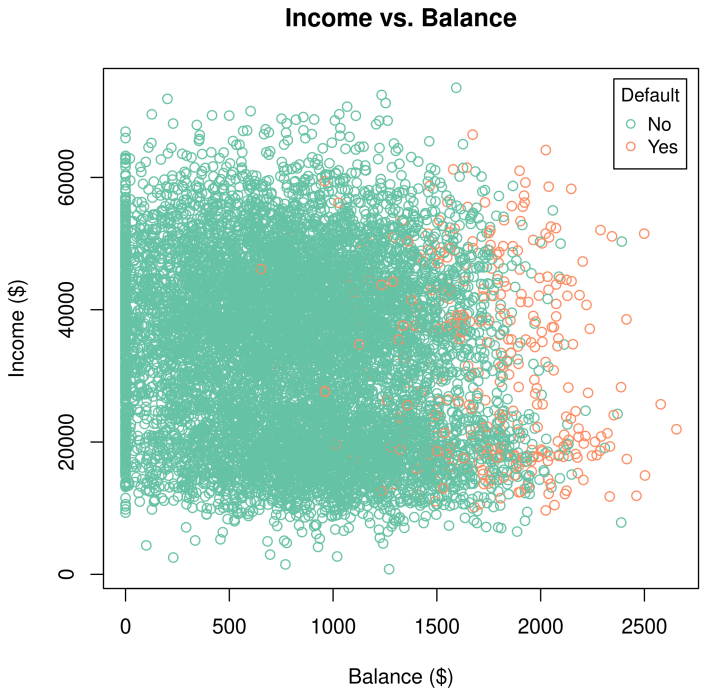

Figure 1: Balance vs. income (Default dataset)

(a) Estimate a logistic regression model to predict the probability of defaulting on credit card *[3]* payments (assume `Default` = "Yes" is the positive class). Write down the estimated logistic function and comment on the significance of the parameter estimates.

(b) Assuming a 50% threshold, use the logistic regression model to assign the default status of the *[1]* observations, i.e.:

$$\Pr(\texttt{default} = \texttt{Yes}|\boldsymbol{x}) > 0.5 \tag{3}$$

[3] **(c)** Generate a plot similar to that in Figure 1 and show the decision boundary at the 50% threshold.

(d) What are the overall error rate, sensitivity (recall) and precision of the classifier? *[3]*

(e) Tabulate or plot the confusion matrix. *[2]*

(f) How would you increase the sensitivity of your classifier? Indicate the action you would take and show the new decision boundary in the same plot generated in part (c). *[2]*

(g) Plot the receiver operating characteristics (ROC) curve based on the model estimated in **(a)** and state the area under the curve (AUC). *[2]*

(h) **[Extra Credit]** What would be the shape of the ROC curve if the `default` response did not depend in any way on `balance` and `income`? Why? *[2]*

Problem 2 *LDA (13 pts)*

(a) Suppose we have features $x \in \mathbb{R}^D$, a two-class response with class sizes n_1, n_2 and the target coded as $\{-n/n_1, n/n_2\}$. Show that the LDA rule classifies to class 2 if *[8]*

$$x^\intercal \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^\intercal \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log(n_2/n_1), \tag{4}$$

and class 1 otherwise. (*Hint:* First write the priors π_1 and π_2. Then write the discriminant functions δ_1 and δ_2. Knowing that the LDA classifier assigns an observation to class 2 when $\delta_2 > \delta_1$, expand this condition to obtain (4).)

(b) Estimate a linear discriminant analsysis model to predict `Default` based on `income` and `balance`. Is this a better fit compared to logistic regression? Discuss. *[5]*

Problem 3 *Linear regression (20 pts)*

The goal of this problem is to apply linear regression techniques to analyze **average vehicle ownership per household** in Massachusetts. The data are from the 2010 United States Census, which has been complemented with information from the American Community Survey (conducted between 2005 and 2010).

Part 3.1 *Model exploration*

For this part, you do not have to structure your analyses as listed below, as long as you cover all the required elements. However, provide an organized summary of your analyses afterward.

(a) Based on an exploratory data analysis, propose a model to explain the average number of vehicles per household, using the 321 towns in the training data set in `Massachusetts_Census_Data_training.csv`. (You may consider the techniques of *subset selection*, *ridge regression*, etc.) *[3]*

(b) Briefly describe the thought and exploratory processes that you followed to arrive at this model selection. *[3]*

(c) Provide all the model estimates, their statistical significance, and all the goodness-of-fit indicators of for the model, supported by one or more plots. (You may use a table or simply list the relevant results.) *[3]*

Part 3.2 *Forecasting and analyzing residuals*

You will now evaluate the performance of the models you estimated using the test set in `Massachusetts_Census_Data_test.csv`. We will refer to the model of best fit you earlier chose as $\mathcal{M}$.

[3] **(a)** Using $\mathcal{M}$, predict the average number of vehicles per household in the 30 test towns. Show your results as both a table and a plot.

[3] **(b)** Now, plot and analyze the residuals for all 321 towns using $\mathcal{M}$, indicating if the assumptions of normality and homoscedasticity hold. If not, discuss how these issues may be addressed.

[3] **(c)** Provide confidence intervals of the predicted values using $\mathcal{M}$, and check if the observed average numbers of vehicles per household are within to the intervals.

Problem 4 *Exploration of shrinkage methods (14 pts; 5 pts EC)*

Consider the special case of performing regression *without an intercept* on a design matrix $\boldsymbol{X}$ with N rows (observations) and D columns (features). The following relationships hold:

$$N = D$$
$$x_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{5}$$

[3] **(a)** Show algebraically that the least squares solution is given by:

$$\hat{\boldsymbol{w}}_j = y_j \tag{6}$$

[5] **(b)** The ridge regression estimate is given by:

$$\hat{\boldsymbol{w}}^R = \arg\min_{\boldsymbol{w}} \left[RSS^R(\boldsymbol{w})\right] = \arg\min_{\boldsymbol{w}} \left\{ \sum_{j=1}^{D} (y_j - \boldsymbol{w}_j)^2 + \lambda \sum_{j=1}^{D} \boldsymbol{w}_j^2 \right\} \tag{7}$$

Show algebraically that the ridge solution is:

$$\hat{\boldsymbol{w}}_j^R = \frac{y_j}{1+\lambda} \tag{8}$$

[3] **(c)** Assume that $D = 2$ (2-dimensional problem with 2 observations), $y_1 = 5$, $y_2 = 1$ and $\lambda = 1$. Using these values, create a three-dimensional plot of $RSS^R(\boldsymbol{w})$ as a function of $\boldsymbol{w}$. Include corresponding contour plots for clarity. Confirm that the ridge solution is indeed given by (8).

[3] **(d)** The lasso estimate is given by:

$$\hat{\boldsymbol{w}}^L = \arg\min_{\boldsymbol{w}} \left[RSS^L(\boldsymbol{w})\right] = \arg\min_{\boldsymbol{w}} \left\{ \sum_{j=1}^{D} (y_j - \boldsymbol{w}_j)^2 + \lambda \sum_{j=1}^{D} |\boldsymbol{w}_j| \right\} \tag{9}$$

The lasso solution is given by:

$$\hat{\boldsymbol{w}}_j^L = \begin{cases} y_j - \frac{\lambda}{2}, & y_j > \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2}, & y_j < -\frac{\lambda}{2} \\ 0, & |y_j| \leq \frac{\lambda}{2} \end{cases} \tag{10}$$

Using the same assumptions as in part (c), create a three-dimensional plot of $RSS^L(\boldsymbol{w})$ as a function of $\boldsymbol{w}$. Include corresponding contour plots for clarity. Confirm that the lasso solution is indeed given by (10).

Problem 5 *Comparing subset selection methods, ridge regression and the lasso (14 pts)*

From the Boston average monthly temperature dataset we considered an exhaustive enumeration approach in selecting a subset of predictors (12 lag variables). The dataset is in the file: `data/boston_monthly_avg_temps_1978_2019.csv`.

(a) Now, apply *any* of these three methods: best subset selection, forward stepwise selection and backward stepwise selection to estimate a model to predict monthly average temperature based on the lag variables in the training set. Briefly describe the quality of your estimate and show the equation of the best estimated model. Note the metric(s) you employ in selecting your model for each of the three cases. Interpret the coefficients. Compute the prediction error using the test set. *[5]*

(b) Implement ridge or lasso regression to predict monthly average temperature. Describe the quality of the fit and compute the test error. Discuss how you selected the optimal hyperparameter λ. Include relevant diagnostic plots. Briefly describe the quality of your estimate and show the equation of the best estimated model. Compute the prediction error using the test set. *[5]*

(c) Summarize the performance of the two models in a table, indicating the R^2, MSE, among other performance metrics of your choice. Create a plot comparing the predicted values from both models compared to the observed. Briefly discuss your observations. *[4]*

Problem 6 *Generalized additive modeling (12 pts; 4 pts EC)*

Read the paper "Estimating $PM_{2.5}$ Concentrations in Xi'an CityUsing a Generalized Additive Model with Multi-Source Monitoring Data" (Song et al., 2015).

The goal of this problem is to estimate a GAM model to explain the 2013 concentration of particulate matter ($PM_{2.5}$) in Xi'an City, China, based on the presence of other atmospheric pollutants and the weather. The form of the model you are to estimate is given in by equation 2 of the paper (page 14).

(a) Estimate the GAM model described in the paper (reserving a portion of the dataset for performance assessment purposes). The authors state that the "model explains 69.50% of the total deviance in the $PM_{2.5}$" data. Do your results match this performance? Explain why, if not. Plot the performance of your model on the test set (e.g. similar to Fig. 11b in the paper). Report key statistics from your model (e.g. MSE). *[12]*

EC See if you can improve the results reported in the paper by enhancing the model. *[4]*