Assignment 1: Regression

Innopolis University

Machine Learning Fall 2021 - Bachelors

1 General Instructions

In this assignment, you will solve several tasks applying regression. For the first one, you need to try different regression models to fill the missing values in each feature with a regression model for each of them. For the second task, you will explore a dataset of email spam and train logistic regression to find junk email, and finally, conclude features that influence the decision the most.

You are required to submit your solutions via Moodle as a single zip file. The zip archive should contain a single ipynb, and a single PDF for the theoretical parts 2.2 and 3.2. Please, put your name and email at Innopolis.university as the first line in the notebook.

Source code should be clean, easy to read, and well documented. Bonus points may be awarded for elegant solutions. However, these bonus points will only be able to cancel the effect of penalties.

Do not just copy and paste solutions from the Internet. You are allowed to collaborate on general ideas with other students as well as consult books and Internet resources. However, be sure to credit the sources you use and type all the code, documentation by yourself.

2 Linear/Polynomial Regression

2.1 Practical Task 1 [25%]

In this task you are going to make data imputation. You are to fill all the missing values in the dataset. That is an important step towards solving any ML problem because missing data decrease accuracy of your model. You have a task1_dataset.csv file where a large percentage of data is lost. Also there is a ground truth dataset task1_dataset_full.csv. There are 4 columns: datetime and 3 numerical features. The 3 numerical features are independent from each other. So you should not use one feature to estimate values of another. Fit models with datetime as X and feature as Y. Your task:

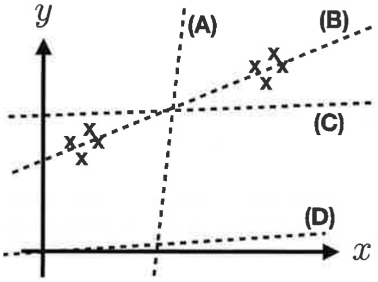
- Preprocess and visualize the dataset: [20%]
  - Encode datetime column with integer values from 0 to len(dataset). It will be easier to do visualization and model training.
  - Plot all features of the dataset (on separate plots). Use matplotlib.pyplot for that.
- Use different regression models with different degrees from this interval $[1, 10]$ to predict missing values and fill the gaps (provide imputation). Don't use imputation libraries. [50%]
- Plot change of MSE for each degree for all features. (MSE between imputed dataset and ground truth one). [20%]
- After experiments with regression models, report best regression degree for each feature. Explain this result: write your ideas on why these degrees best describe particular feature. [10%]

2.2 Theoretical Question On Ridge Regression [20%]

In the case of 1D data, the ridge regression estimator produces:

$$(\hat{\theta}, \hat{\theta_0}) = \underset{\theta,\theta_0}{argmin} \sum_{t=1}^{n} (y_t - \theta x_t - \theta_0)^2 + \lambda\theta^2 \tag{1}$$

for some $\lambda > 0$, and then makes predictions of the form $\hat{y} = \hat{\theta} x + \hat{\theta_0}$. In this question, we consider a 1D data set with 8 points $(x_t, y_t)_{t=1}^{8}$, marked with 'X - a cross' in the following figure:



The four dashed lines, labeled (A), (B), (C), and (D), each correspond to a linear prediction rule: Given a new x, each model predicts y to be the corresponding point on the line. For each of (A), (B), (C), and (D), indicate which of the following statements is the most appropriate:

- **High λ.** The prediction rule could be produced by ridge regression with a high λ value;
- **Low λ.** The prediction rule could be produced by ridge regression with a low λ value;
- **Neither.** The prediction rule could not be produced by ridge regression.

Choose only one of these options for each of (A), (B), (C), and (D), and include a brief explanation (ideally only 1-2 sentences with minimal use of equations) for your choice.

3 Logistic Regression

3.1 Practical Task 2 [35%]

In this task you're going to build a model for loan applicant approval. The goal is to classify loan applicants into one of two categories, good or bad. The dataset contains 1000 records of bank information of applicants represented with 20 attributes (7 numerical, 13 categorical). A full description of the dataset is attached.

Your task:

1. Preprocess and visualize the dataset:
   - Transform all categorical values into numerical values. You are free to apply any ways of handling categorical data and missing values.
   - Scale features if necessary. Explain why did you decide to scale or not.
   - Visualize the dataset in two dimensions. Dimension reduction methods such as PCA can be used.
   - Using pandas built-in function, plot the correlation matrix. Answer questions: Are there highly co-related features in the dataset? Is it a problem for regression task?
2. Split the dataset into train(70%) and test set(30%).
3. Apply logistic regression using linear and non linear function. Use different polynomial models with different degree (range from 1 to 10) and select the model which performs the best in terms of bias-variance. Highlight which model (degree) underfit and overfit the data *hint: plot the train and test error for each model.*

4. Now that you have the best degree for your model, you will make it even better by fine-tuning its hyperparameters. Use GridSearchCV to find the best hyperparameters. Try different variations with penalty ['$l1$', '$l2$'], type of solver: ['liblinear', 'lbfgs'], and regularization strength: $np.logspace(4, 4, 20)$.
5. Using your best model, compare the accuracy of predictions across male and female applicants e.i, split the test set into two groups (Male and female) compute the accuracy on each groups using the test set, plot and compare them. What conclusion can you draw and what could be the source of your observation?

3.2 On Regularization in Logistic Regression [20%]

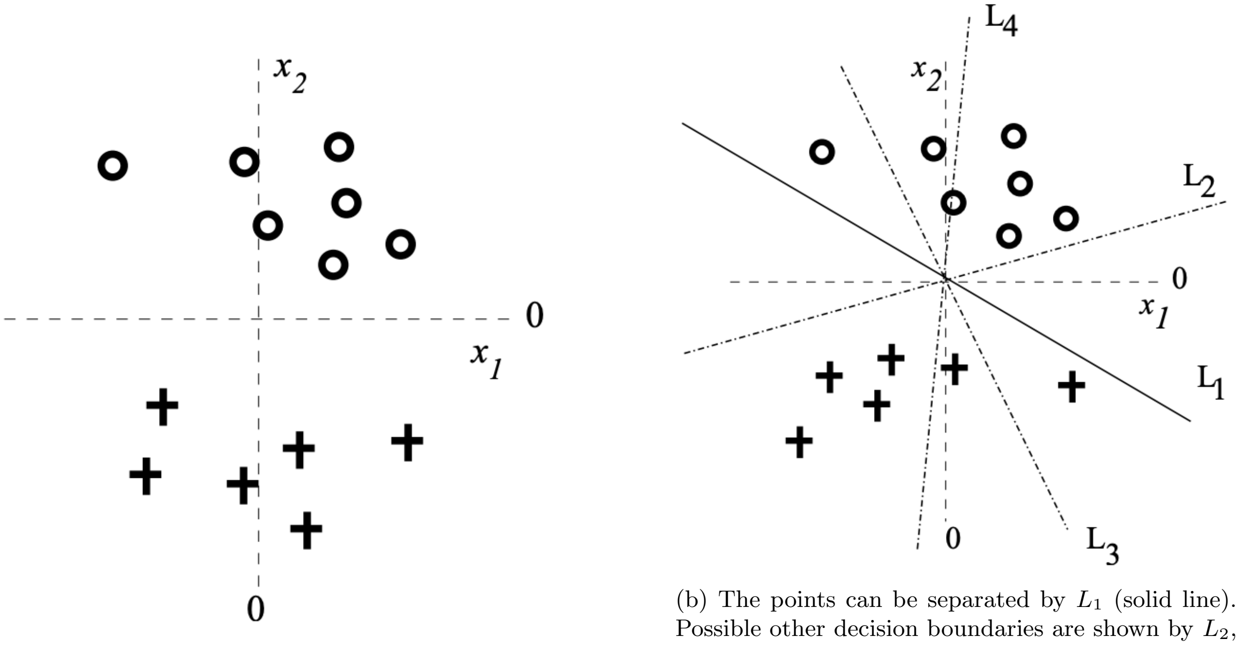
In this problem we will refer to the binary classification task depicted in Figure 1a, which we attempt to solve with the logistic regression model

$$\hat{P}(y = 1|x; \theta_1, \theta_2) = g(\theta_1 x_1 + \theta_2 x_2) = \frac{1}{1 + exp(-\theta_1 x_1 - \theta_2 x_2)}$$

(for simplicity we do not use the intercept parameter θ_0). The training data can be separated with zero training error – see line L_1 in Figure 1b, for instance, which is the line obtained with no regularization. Consider a regularization approach where we try to maximize

$$\sum_{i=1}^{n} log\ p(y_i|x; \theta_1, \theta_2) - \frac{C}{2}\theta_2^2$$

for large C. Note that only θ_2 is penalized. Recall also that line L_1 in the figure corresponds to $C = 0$. We would like to know which of the four lines in Figure 1b could arise as a result of such regularization. For each potential boundary L_2, L_3, and L_4, determine whether it can result from regularizing θ_2. If not, explain briefly why not.



(a) The 2D dataset used in Problem (3.2).

(b) The points can be separated by L_1 (solid line). Possible other decision boundaries are shown by L_2, L_3, L_4.

Figure 1: Binary classification.

4 Notes

- Cheating is a serious academic offense and will be strictly treated for all parties involved. So delivering nothing is always better than delivering a copy.
- Late assignments will not be accepted and will receive **ZERO** mark.
- Code cleanness and style are assessed. So maybe you want to take a look at our references: Link 1 and Link 2.
- Organize your notebook appropriately. Divide it into sections and cells with clear titles for each task and subtask.