BAX 452 HW 2

Course Instructor : Dr. Rahul Makhijani

Submit the assignment on canvas. The deadline is Sunday Jan 26 11:59 pm
Please note: Late homework will not be accepted.

Cross Validation 30 points

In class we saw that cross validation is a way to estimate the generalization error. In this exercise we will implement the cross validation method.

We want to perform a regression of the form $Y \sim \sum_{k=1}^{d} X^k$ and use cross validation as a method to identify the optimal d.

1. We first construct our dataset in the following way: -
   Sample a vector $X \in R^n$ where each $X_i \in U[0,1]$. Each sample point X_i is sampled from the uniform distribution. Construct Y from X using the following equation

   $$Y = 3 \times X^5 + 2 \times X^2 + \epsilon$$

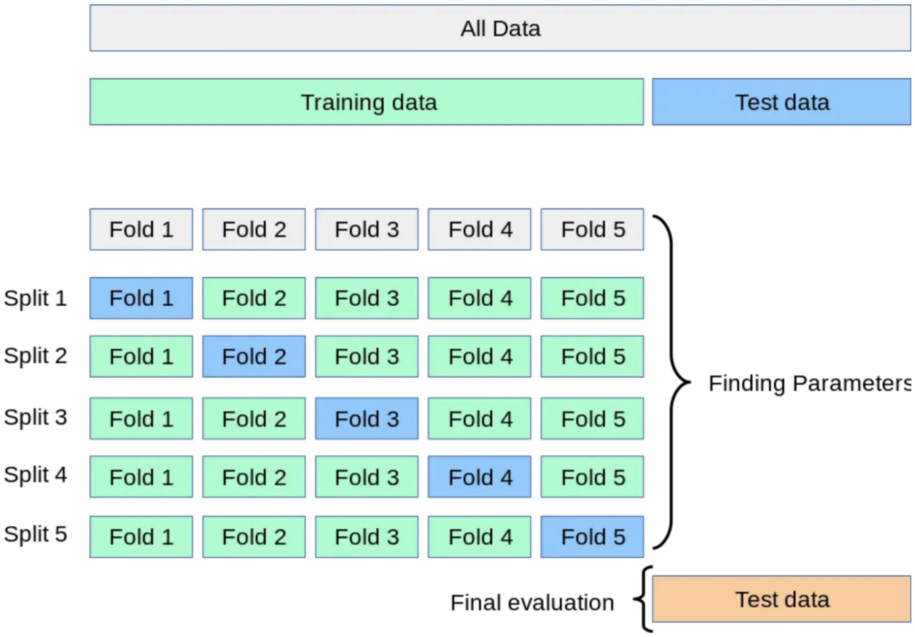
   where $\epsilon \in R^n$. Each ϵ_i is sampled independently from the N(0,0.5) (normal distribution)

   Choose $n = 10000$.

2. Split the 10000 points into a 80% training and 20% test split. Use a seed before randomizing to replicate results.

3. Split the training set into 5 parts and use the five folds to choose the optimal d. The loss function you would implement is the MSE error. You want to estimate the MSE error on each fold for a model that has been trained on the remaining 4 folds. The cross validation (CV) error for the training set would be the average MSE across all five folds. Plot the CV error as a function of d for d $\in [1, 2, \ldots, 10]$

   Note: Code the cross validation function from scratch - do not use any package.



The figure below gives a diagrammatic view of what we want to do.

4. In this subpart, use the entire training set for training the models. Compute the performance of the 10 models on the test set. Plot the test MSE and training MSE as a function of d. Comment on your observations.

Bias Variance Tradeoff 25 points

We also learnt in class about bias variance tradeoff as we increase model complexity. and see the bias variance tradeoff via simulation. For this exercise, we would use the entire training set to train models of the form $Y \sim polynoial(X, d)$. You should compare the bias and variance of 10 different models (for d $\in [1, 2, \ldots, 10]$). For each of the models, compute the bias and variance while predicting the output at a new test point $x = 1.01$.

Here we would look at 1000 datasets instead of just one dataset. Use the same function form to generate the data as in the previous question but now use $n = 100$.

1. For each of the simulated training dataset you generated, train 10 different models (d $\in [1, \ldots, 10]$.) Store and compute the prediction for $x = 1.01$

2. Calculate the bias and variance of the prediction value. Plot the bias and variance as a function of d.

3. Consider the two cases below

   (a) Plot happens to bias and variance if we instead sample from $X_i \in U[0, 10]$ instead

   (b) Plot what happens to bias and variance if we instead use test point x = -0.5 ?

   Can you explain why do the plots look like above ? What are the implications ? Can we mitigate any of the issues ?

Regularized parameter derivation 15 pts

Consider the regularized model with $n = p$ (number of predictors same as the number of samples). Consider the special case with X is the diagnol matrix. (i.e. the i^{th} sample point is $(y_i, 0, \ldots, 1, \ldots, 0)$ (1 in the i^{th} position of X) Assume that the intercept term is 0.(y is centered)

- Derive the ridge and lasso parameters w_r and w_l in terms of y and the penalty parameter λ.

- In case of lasso you can write the solution in terms of least square solution and the penalty parameter λ.

Grading split

- Ridge parameter derivation - 5pts

- Lasso parameter derivation - 10 pts)

Credit Card Fraud Detection for Imbalanced Dataset (30 Points)

Dataset: Use the *Kaggle Credit Card Fraud Detection dataset* (`https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`). This dataset contains transactions made by European cardholders in September 2013, with 284,807 transactions where only 492 are fraudulent (highly imbalanced).

Task:

You are tasked with building a machine learning model to classify transactions as fraudulent or legitimate using the provided dataset. Your solution should address the challenges of imbalanced data and consider business metrics for evaluating the model's performance.

1. **Data Understanding and Preprocessing (5 Points)**
    - Perform exploratory data analysis (EDA) to understand the dataset. Discuss the class imbalance and its implications on model performance.
    - Propose preprocessing steps, including handling the imbalanced dataset and feature scaling.
2. **Model Building and Training (10 Points)**
    - Train a Logistic Regression model to perform classification. Use a 80:20 split for splitting the training and test data.
    - Check if regularization is helpful in this case or not.
3. **Evaluation and Metrics (15 Points)**
    - Evaluate your model using business-relevant metrics such as precision, recall, F1-score, and the ROC-AUC score.
    - Discuss why metrics like accuracy might not be appropriate in this scenario.
    - Calculate the expected financial loss for a given confusion matrix, considering the following costs:
        - Missing a fraudulent transaction (False Negative): $500

    - Incorrectly flagging a legitimate transaction as fraudulent (False Positive): $10
- Optimize the model to minimize the total financial loss.

Deliverables:

- A written report discussing the data exploration, preprocessing, cross validation and modeling process.
- Visualizations of metrics and results.
- Calculation of financial loss based on the confusion matrix.