Indian Institute of Technology Madras

ID5055 Foundations of Machine learning

Assignment III

Due date: 14^{th} October 2023

Instruction

1. Assignment shall be submitted on the due date. Late submissions will not be entertained. If you cannot submit the assignment due to some reasons, please contact the instructor by email.
2. All the assignments must be the student's own work. The students are encouraged to collaborate or consult friends. In the case of collaborative work, please write every student's name on the submitted solution.
3. If you find the solution in the book or article or on the website, please indicate the reference in the solutions.

Problems

1. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the i^{th} fitted value takes the form

$$\hat{y}_i = x_i \hat{\beta} \tag{1}$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i \right) / \left(\sum_{i=1}^{n} x_i^2 \right) \tag{2}$$

Show that we can write

$$\hat{y}_i = \sum_{i=1}^{n} a_i y_i. \tag{3}$$

What is the value of a_i?

2. The ridge regression objective function is defined as

$$\begin{aligned} J(\theta) &= \frac{1}{2} \sum_{i=1}^{m} (\beta^T \mathbf{x}_i - y_i)^2 + \frac{\lambda}{2} ||\beta||^2 \\ &= \frac{1}{2} ||\mathbf{X}\beta - \mathbf{y}||_2^2 + \frac{\lambda}{2} ||\beta||^2. \end{aligned}$$

Find the closed form expression for the value of β which minimizes the ridge regression objective function.

3. You are given a design matrix X (whose i^{th} row is sample point x_i^T) and an n-vector of labels $y \triangleq [y1 \dots y_n]^T$. For simplicity, assume $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$. Do not add a fictitious dimension/bias term. For input 0, the output is always 0. Let x_{*i} denote the i_{th} column of $\mathbf{X}$.

(a) Show that the cost function for L_1-regularized least squares, $J_1(\beta) \triangleq ||X\beta - y||^2 + \lambda||\beta||_1$ (where $\lambda > 0$), can be rewritten as $J_1(\beta) = ||y||^2 + \sum_{i=1}^{d} f(x_{*i}, \beta_i)$ where $f(\cdot, \cdot)$ is a suitable function whose first argument is a vector and second argument is a scalar.

(b) Using your solution in the previous question 3a, derive the necessary conditions for the i^{th} component of the optimizer β^* of $J_i(\cdot)$ to satisfy each of these three properties: $\beta_i^* > 0$, $\beta_i^* = 0$ and $\beta_i^* < 0$.

(c) For the optimizer $\beta^{\#}$ of the L_2-regularized least squares cost function $J_2(\beta) \triangleq ||X\beta - y||^2 + \lambda||\beta||^2$ where, $\lambda > 0$, derive a necessary and sufficient condition for $\beta_i^{\#} = 0$, where $\beta_i^{\#}$ is the i^{th} component of $\beta^{\#}$.

(d) A vector is called *sparse* if most of its components are 0. From your solution to part 3b and 3c, which of β^* and $\beta^{\#}$ is more likely to be sparse? Why?

4. You are the visionary owner of "HealthPlus Insurance", a prominent health insurance company dedicated to improving the healthcare financing landscape. Your company's long-term success hinges on accurate predictions of medical expenses, allowing you to set competitive premiums while ensuring profitability. Medical expenses are influenced by various factors, including age, smoking habits, and obesity. As the owner, you are personally assuming the role of Chief Data Scientist, responsible for analysing the provided dataset ("insurance.csv") and developing a model to estimate average medical care expenses for different population segments.

   (a) Calculate and interpret the correlation matrix to understand relationships among features.

   (b) Create a scatterplot matrix to visualize relationships among features. Explain the insights they can gain from these visualizations.

   (c) Perform data preprocessing and cleaning, which involves addressing missing values and handling categorical features, followed by conducting a train-test split of the data.

   (d) Implementing and training the linear regression model (apply Ridge and Lasso regression techniques) using appropriate Python libraries.

   (e) Evaluate the model's performance by calculating relevant metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. Additionally, interpret the model's coefficients and discuss how various features impact predictions of medical expenses.

5. In Multiple Linear Regression, the normal equation solution was obtained by minimizing the sum of squares error. Show that the Maximum Likelihood method is the in essence the same as minimizing the sum of squares error and thus show that the Maximum Likelihood estimate for the matrix of the coefficients (θ_{ML}) is the same as that obtained via solving the normal equation (Hint: Use the assumption that the additive noise term is Gaussian)

6. Using K-Means clustering for image compression: Image compression enables us store enormous amounts of data while using lesser disk space while retaining significant aspects of the image which will be needed for analysis. We can apply the K-Means algorithm to an image where the parameter K is the palette of the colours that we have in the final image.

   (a) Use the K-Means algorithm to apply compression on a test image. Visualize the results obtained by using powers of 2 less than 2048 as the value for K

   (b) Decide on an appropriate value for K (Hint: Use an elbow plot to justify your choice)

   (c) Is the compression obtained lossy or lossless? What is the effect of varying the value of K in terms of overfitting or underfitting the data?

   A color image is represented by a matrix of dimensions (w,h,c) where w, h, c stand for width, height, and number of color channels which in our test image is three, (for example RGB: Red, Green and Blue). These three colors can be treated as three features and each pixel can be treated as separate datapoints on which we apply the K-Means clustering.

7. Visualize the difference in the clusters obtained through applying Hierarchical Clustering on Online Retail data using different linkage methods. Plot the dendograms for the three linkage methods(single, complete and average). Use the number of clusters as 3 for producing the visualizations. Provide a brief explanation for the use cases of each kind of linkage.

8. Imagine that you are running "MedGenius Solutions", a startup providing innovative healthcare solutions, and you aim to develop a Proof of Concept (PoC) using a toy dataset. Your goal is to secure a project and budget from leading clients in the healthcare industry. To achieve this, you need to demonstrate the capabilities of spectral clustering in gene expression analysis. You are tasked with developing a PoC for using spectral clustering in gene expression analysis. Gene expression data, which records the activity levels of genes across different samples, is provided in tabular form. Your goal is to

   (a) Implement spectral clustering using Python and scikit-learn to identify clusters of co-expressed genes within the dataset.

   (b) Create visualizations for the true clusters based on the information in the 3rd column of the dataset.

   (c) Evaluate and provide insights on the outcomes, including a comprehensive report on performance metrics such as Adjusted Rand Index, Adjusted Mutual Information, and Silhouette Score.

9. Imagine you are conducting a data analysis project and want to use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to cluster data points. You are tasked with implementing DBSCAN, a density-based clustering algorithm, on a dummy toy dataset. The aim is to demonstrate the clustering capabilities of DBSCAN and assess its sensitivity to parameter choices.

   (a) Generate a dummy toy dataset with varying densities and shapes. Set the `eps` (Epsilon) and `min_samples` (MinPts) parameters, and then fit DBSCAN to the generated dataset.

   Hint: You can use functions like `make_blobs` or `make_moons` from scikit-learn to create a synthetic dataset.

   (b) Experiment with each combination of `eps` and `min_samples` (consider at least 3 values of each) for these parameters. Report the values of the performance metrics to evaluate DBSCAN's sensitivity to parameter choices.

   (c) Visualize the clustering results using a scatter plot, where each cluster is assigned a different color. Additionally, use a different marker shape for noise points.

   (d) Calculate and report the following performance metrics: Silhouette Score, Adjusted Rand Index, Adjusted Mutual Information.