Due Date: Wednesday, September 23, 2022, 11:59 PM

Abstract

In this assignment, you will learn about linear regression and classification. The questions below can be answered with the help of the book. Show all calculations done step-by-step. In problem 3, you need not worry about being "mathematically" correct. Please follow the same notation as in the book. You are allowed to work on teams of up to three people. If working as a team, every member of the team should submit the (same) report. All reports and supplemental material must be zipped as `team#_Assignment#.zip` and uploaded on Blackboard.

Problem 1. *Explain the relationship between shrinkage methods and feature selection. Disregarding computational costs, why would ridge regression be favored over any of the feature selection methods? Show your reasoning mathematically.*

Solution:
Relationship between shrinkage methods and feature selection:

1. Shrinkage methods and feature selection help in improving the 'least square' estimates which aren't 'interpretable' and have 'prediction errors'. 'Feature Selection' helps in producing interpretable models but being a discrete process, it can't control the variance which results in prediction inaccuracy. With 'Shrinkage Methods' being continuous there'll be no too much variance, helping in getting accurate predictions. Feature selection applies 'hard thresholding' on the least squares estimates by dropping all variables having coefficients $<$ threshold value, whereas ridge regression performs 'soft thresholding' by doing proportional shrinkage on the coefficients.

2. The penalty parameters for ridge regression vary continuously whereas for feature selection it takes distinct steps to reach the least squares estimate. While converging to the least squares estimate, ridge regression shrinks the coefficient together, whereas the feature selections overshoots and then backtrack.

3. Ridge regression shrinks more towards low-variance directions, unlike feature selection, which makes Ridge regression more stable when compared to feature selection. Hence in-order to reduce the prediction error, ridge regression is preferred over feature selection. In conclusion, Ridge regression is opted as it shrinks smoothly rather than in discrete steps(like feature selection).

Mathematical Reasoning:

$$\text{RSS(Residual Sum of Squares)} = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

Ridge regression is represented by: RSS + (λ + (m $*$ m))

Let's assume there is one feature and one target variable in a plane with 3 different slope(m) (0, 0.2, 0.4) and we are going to substitute λ and b(slope) values to the ridge regression equation above:

1. When λ is zero (No penalty):
1.1 Slope 0: RSS + (0*(0*0)) = RSS
1.2 Slope 0.2: RSS + (0*(0.2*0.2)) = RSS
1.3 Slope 0.4: RSS + (0*(0.4*0.4)) = RSS

2. When λ is 10 (Imposing penalty 10):
2.1 Slope 0: RSS + (10 * (0*0)) = RSS
2.2 Slope 0.2: RSS + (10*(0.2*0.2)) = RSS + 0.4
2.3 Slope 0.4: RSS + (10*(0.4*0.4)) = RSS + 1.6

3. When λ is 40 (Imposing penalty 40):
3.1 Slope 0: RSS + (40 * (0*0)) = RSS
3.2 Slope 0.2: RSS + (40*(0.2*0.2)) = RSS + 1.6
3.3 Slope 0.4: RSS + (40*(0.4*0.4)) = RSS + 6.4

When pointing the above points on the graph, we can observe, all the sum of residuals start at same point at all the slope values. Also, RSS value decreases with increasing slopes. But, when λ increases, the optimal value of the slope moves closer to zero but never becomes zero. By, this we can conclude saying ridge regression shrinks the optimal value of the slope but never shrinks till zero. It means, ridge regression never nullifies the effect of the less competent features. Whereas the feature selection models remove them completely. This creates a problem when that feature combines with other feature and correlates with target variable together.

Problem 2. *State the three assumptions of linear regression. Then, compute the column rank of matrix A:*

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 3 & 4 & 7 \\ 4 & 5 & 9 \end{bmatrix}$$

Based on your calculated rank of the matrix A, how does this affect linear regression, and what are the possible solutions one might employ, if any?

Solution: Assumptions of Linear regression are:
1. Linearity: A linear relationship must be there between independent variables and dependent variable.
2. Normality: Residuals are independent and normally distributed.
3. Homoscedasticity: Homogeneity of variances, equal or similar variances for various independent variables.
4. No Multicollinarity: There should not be any high correlation between independent vari-

ables.

The calculated rank for the given matrix A is 2. As the number of columns present in A is 3 and rank obtained is 2, it is not considered as a Full Rank.

To explain further; the equation of the linear regression is expressed as:
f(x) = (b + mx) + e; where b is the y-intercept, m is the slope and e is the error/residual.

The matrix equation is expressed as: y = XA + E. To get a unique solution for A, the rank of the matrix should be full rank else we get multiple solution and leads to linear dependency problem. To our understanding, rank deficiency leads to multicollinearity problem. To overcome the multicollinearity, the possible solution is to remove the highly correlated variables.

Problem 3. *Design your own feature selection algorithm (conceptually). Explain your algorithm's advantages and disadvantages compared to the three feature selection methods covered in class. You can focus on a specific type of data if you wish.*

Solution: Our feature selection model calculates 4 scores for every features.
1. Calculates the number of missing values in the feature.
2. Calculates the correlation of the feature with the target variable.
3. Calculates the feature variation intensity. i.e., repetation of same value in the features.
4. Calculates the multicollinearity using VIF.

We consider 25% from each above calculated scores and sort the features in descending order.

We now, remove the features with highest percentage from the full model and compare the r-squared value with the full model's r-squared value. If, the value increases, we don't consider that feature anymore. And repeat the same procedure with other features in the sorted list.

We stop the procedure when r-square value drops from the previous model and consider the obtained model as optimal model.

Advantages:
1. It is computationally not expensive as feature selection models because we don't do any feature combinations.
2. We can implement this method on any size data. Be it large or small. This method works efficiently.

Disadvantages:
1. We cannot implement any methods to handle the missing data because missing data is one of the criteria for feature selection.
2. This selection may suffer a bit with few features like BMI because most of the rows have same value in it(low variance)

Problem 4. *One method of solving the logistic regression equations is to use the Newton-Raphson algorithm; how would a large number of features p > 1000 affect the algorithm's performance (hint: Hessian matrix)?*

Solution: 'Newton Raphson' algorithm with the help of second order derivate square Hessian Matrix logistic regression model is fit. But it's performance is effected with the number of features(p) involved. It is a iteratively reweighted least squares algorithm and based on the number of features(p) value, it computes weighted least squares for each entry in the matrix resulting in an adjusted response. With p value > 1000, it results in more than 1 million entries in the matrix(p * p) and performing weighted least square on each entry will take a lot of time to compute, memory and sometimes it can result in overfitting of the model there by decreasing the algorithm's performance.

Problem 5. *We are told that logistic regression is more general than linear discriminant analysis. Identify the differentiating reason between them and explain how this affects the performance of linear discriminant analysis; what effect do outliers carry on both methods?*

Solution: In **Linear Discriminant Analysis**, the logits of the odds ratio is linear, because of the assumptions of '**Gaussian Distribution**' and **common covariance matrix**. **Logistic Regression** too has linear logits, but both(LDA and LR) are not same, the main differentiating reason is how the 'linear coefficients' are estimated and logistic regression makes less assumptions.

$$\mathbf{Pr(X,\ G = k) = Pr(X) * Pr(G{=}k|X)}$$

P(X) - marginal density function

Pr(G=k—X) - Posterior Probability is same for both LR and LDA.

In logistic regression we fit the parameters by by **maximizing the conditional likelihood**, here '**marginal density**' has no assumptions. **Marginal density** though totally ignored is estimated in complete non-parametric way with no constraints by dividing every result by 'N'.

Parameters are fit by full log-likelihood maximization in linear discriminant analysis. The marginal density plays a key role and is not ignored, it is a mixture density and is estimated in a parametric way. Here by assuming the distribution is Gaussian we can estimate them more efficiently with lower variance. But by ignoring the 'marginal density' may **make the model to perform only with 70% accuracy**.

Outliers play a role in computing common covariance matrices is an assumption by LDA, but that's not the case with LR. **Not having this assumption makes LR more robust to outliers when compared to LDA.**

Consequences caused by the assumptions(**Gaussian Distribution** and **common covariance matrices**) make LDA less robust when compared to LR's performance. Because the assumptions can never be correct and hence LR is a safer, more robust than LDA model because of less assumptions.

Problem 6. *You will be given a synthetic dataset; your task is to implement linear regression on the dataset. First, you must determine whether the given data satisfies the three assumptions of linear regression. Next, you will implement linear regression using the provided code based on your findings. Finally, you will produce the mean squared error for train and test subsets and final coefficient values.*

Solution: Collab notebook link along with analysis: Question 6

Analysis: Three Assumptions of Linear Regression are: Multivariate Normal, No multicollinearity, Homoscedasticity

The given synthetic dataset:

Is normally distributed. We checked the normality using Probability plot and Normal Distribution Curve. In both the checks, we found the given data is satisfying the multivariate normal assumption.

Has no multicollinearity. We checked the multicollinearity variance inflation factor and also confirmed the results by the heatmap. None of the feature's VIF value is above 5. So, we concluded the dataset has no multicollinearity.

Satisfies Homoscedasticity assumption.A close observation of the obtained plot shows that variance of residual terms is distributed evenly for high and low fitted values. Assumptions between predicted values and residuals is verified using scatter plots.

Mean squared error for test set: 9.15040786980671

Mean squared error for training set: 9.477015206695722

MSE value for training and test sets has no significant difference.

Final coefficient values: [69.79975592, -9.3443124, 102.70481272, -6.36292528, 5.87309176, -1.69175087, 42.87669847, -4.07934221, 32.55576259, 32.45166987, 55.41335973, -6.7845724, 4.74741502, 68.70214188, 10.49642582]

We can observe true and learned coefficients and both are almost same as both the coefficients are from the same linear model.

Analysis of Summary table of OLS Statsmodel:

When checked, R-squared and Adjusted R-squared values, there's no significant difference. Therefore, we can conclude all features are relevant to the final model and to fit the model no need to omit any features.

To calculate the overall significance of a regression model F-Statistic is used. The f-statistic probability value is close to 0 in our case, we can conclude our model is better and the target variable has linear relationship with all the features in the model.

Problem 7. *You will be given a different synthetic dataset; your task is to implement a logistic regression model with two other regularization methods. Finally, report your findings, including train and test mean squared errors and final coefficient values. Analyze the difference in coefficient values between the two regularization methods and describe what you understand.*

Solution: Collab notebook link along with analysis: Question 7

References

[1] https://hastie.su.domains/Papers/ESLII.pdf
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/
https://www.kdnuggets.com/2021/12/alternative-feature-selection-methods-machine-learning.html
https://jeheonpark93.medium.com/ml-subset-selection-shrinkage-methods-74d6e99ecaaf
https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning
https://medium.com/geekculture/essential-guide-to-handle-outliers-for-your-logistic-regression-model-63c97690a84d
https://thelaziestprogrammer.com/sharrington/math-of-machine-learning/solving-logreg-newtons-methodhttps://medium.com/analytics-vidhya/create-your-own-coefficient-plot-function-in-python-aadb9fe27a77
https://www.geeksforgeeks.org/normal-probability-plot/
https://www.statology.org/seaborn-normal-distribution/
https://medium.com/analytics-vidhya/how-to-check-for-assumptions-in-a-linear-regression-a68116aef88a https://scikit-learn.org/stable/modules/generated/sklearn.linear$_m$odel.Logisti
$https://realpython.com/logistic-regression-python/$
$https://www.projectpro.io/recipes/perform-logistic-regression-sklearn$
$https://scipy-lectures.org/packages/scikit-learn/index.html$
$https://www.sharpsightlabs.com/blog/sklearn-predict/$
$https://towardsdatascience.com/fit-vs-predict-vs-fit-predict-in-python-scikit-learn-f15a34a8d39f$
$https://scikit-learn.org/stable/modules/model_evaluation.html$
$https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62$