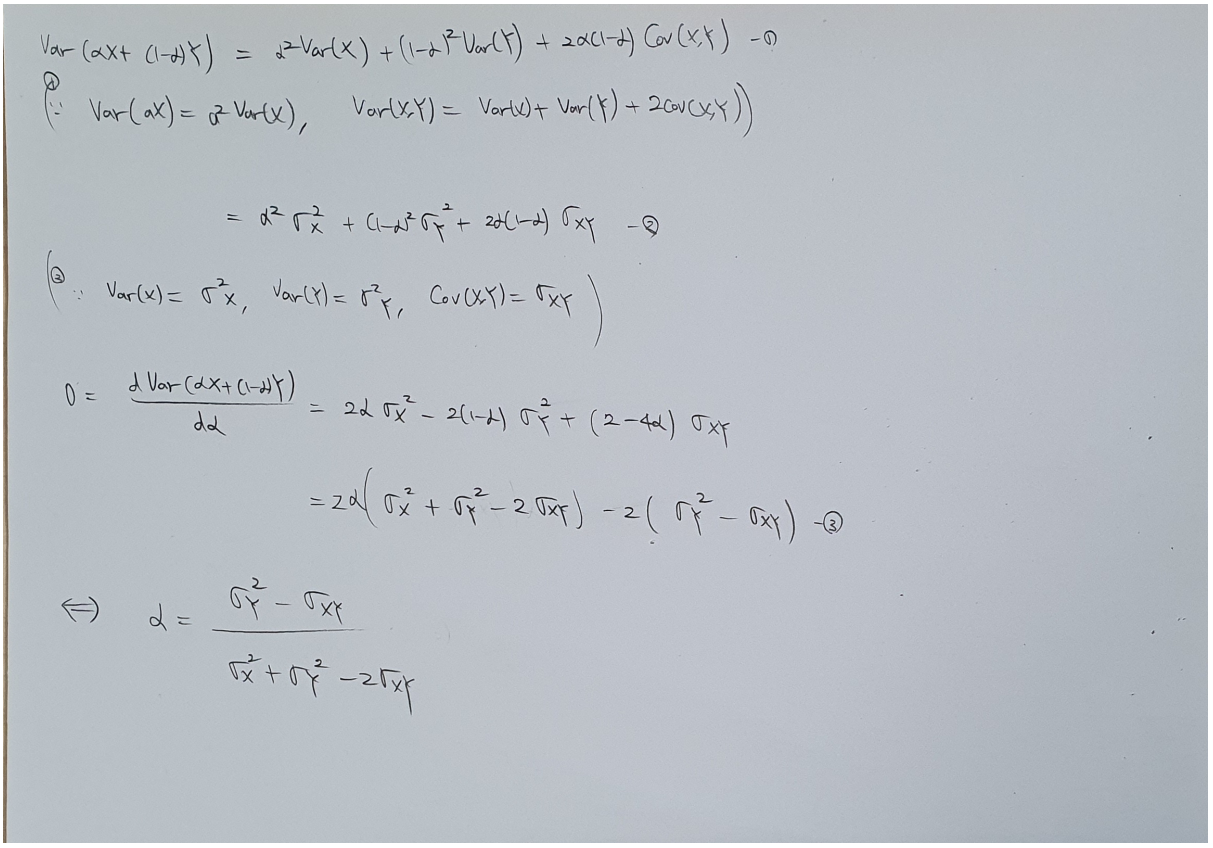
MLDL assignment3

Q1. Prove that α mimimizes Var(αX + (1 − α)Y).

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

$$\sigma_X^2 = Var(X), \sigma_Y^2 = Var(Y), \sigma_{XY} = Cov(X, Y)$$

Sol) Drive that (1) using the property of variance. To find alpha that minimizes the given function, differential by alpha. The following figure explains the whole process.



Q2. Best subset selection vs. Forward stepwise selection

What could be the reason that the fourth model in best subset selection and forward stepwise selection differ as shown in the table below?

Variables	Best subset selection	Forward stepwise selection

One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	**cards, income, student, limit**	**rating, income, student, limit**

Sol) Due to the difference in the selection algorithm, the subset selection sequence will be different. For the Best subset selection, check all possible combinations of variables and choose the best one based on the criterion (ex. AIC, BIC, cross-validation). Thus, we need to check 2^p models. In four variable model using best subset selection, select "'cards', 'income', 'student', 'limit'" is the best case according to the criterion. However, the Forward stepwise selection, starts with an empty model and adds variables based on the criterion (ex. AIC, BIC, cross-validation). Then we need to check 1 + p(p+1)/2 models. In four variable model using forward stepwise selection, "'rating', 'income', 'student', 'limit'" is the best case according to the criterion. Since, the three variable model is "'rating', 'income', 'student'" and adding 'limit' make the best case according to the criterion.

Q3. Subset selection

We perform the best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing 0, 1, 2, · · ·, p predictors. Explain your answers:
(a) Which of the three models with k predictors has the smallest training RSS? [2 pts]

Sol) Performing the best subset selection gets the smallest training RSS. Because for each k predictor, the best subset selection searches all possible cases.

(b) Which of the three models with k predictors has the smallest test RSS? [2 pts]

Sol) In many cases, performing the best subset selection gets the smallest test RSS. For the overfit test model, forward stepwise, or backward stepwise selection might have the smallest test RSS.

(c) True or False [5 pts; 1 pts each]:

i. True. ii. True. iii. False. iv. False. v. False.

i. The predictors in the k-variable model identified forward stepwise are a subset of the pre-

dictors in the (k + 1)-variable model identified by forward stepwise selection.

Sol) True. It is the definition of forward stepwise.

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)-variable model identified by backward stepwise selection.

Sol) True. It is the definition of backward stepwise.

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)-variable model identified by forward stepwise selection.

Sol) False. The selection of predictors does not have to be sorted.

iv. The predictors in the k-variable model identified forward stepwise are a subset of the predictors in the (k + 1)-variable model identified by backward stepwise selection.

Sol) False. The selection of predictors does not have to be sorted.

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

Sol) False. Best subset selection checks all possible cases, the predictors in the k model do not include the k+1 model possible.

Q4. Regression with regularization

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j|\le s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

Sol) Linear regression models regularized with an L2.

(a) As we increase s from 0, the training RSS will:
i. Increase initially, and then eventually start decreasing in an inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

Sol) (iv) Steadily decrease. If the s increases from 0, then each coefficient is also can increase from 0. That means coefficients could be more fit to the train data. Thus, training RSS is steadily decreasing.

(b) Repeat (a) for test RSS.

Sol) (ii) Decrease initially, and then eventually start increasing in a U shape. For the best s value which makes the best model has the lowest test RSS. If the s decreases from the best s value, the test RSS increases. Because the model is underfitting. If the s increases from the best s value, the test RSS is increasing. Because the model is overfitting.

(c) Repeat (a) for variance.

Sol) (iii) Steadily increase. If the s increases from 0, then the model is more flexible. Thus, the variance will be increasing.

(d) Repeat (a) for (squared) bias.

Sol) (iv) Steadily decrease. If the s increases from 0, then the regularization decreases. Thus, the bias will be decreasing.

(e) Repeat (a) for the irreducible error.

Sol) (v) Remain constant. The model cannot be reduced irreducible error, because it was inherent in data.

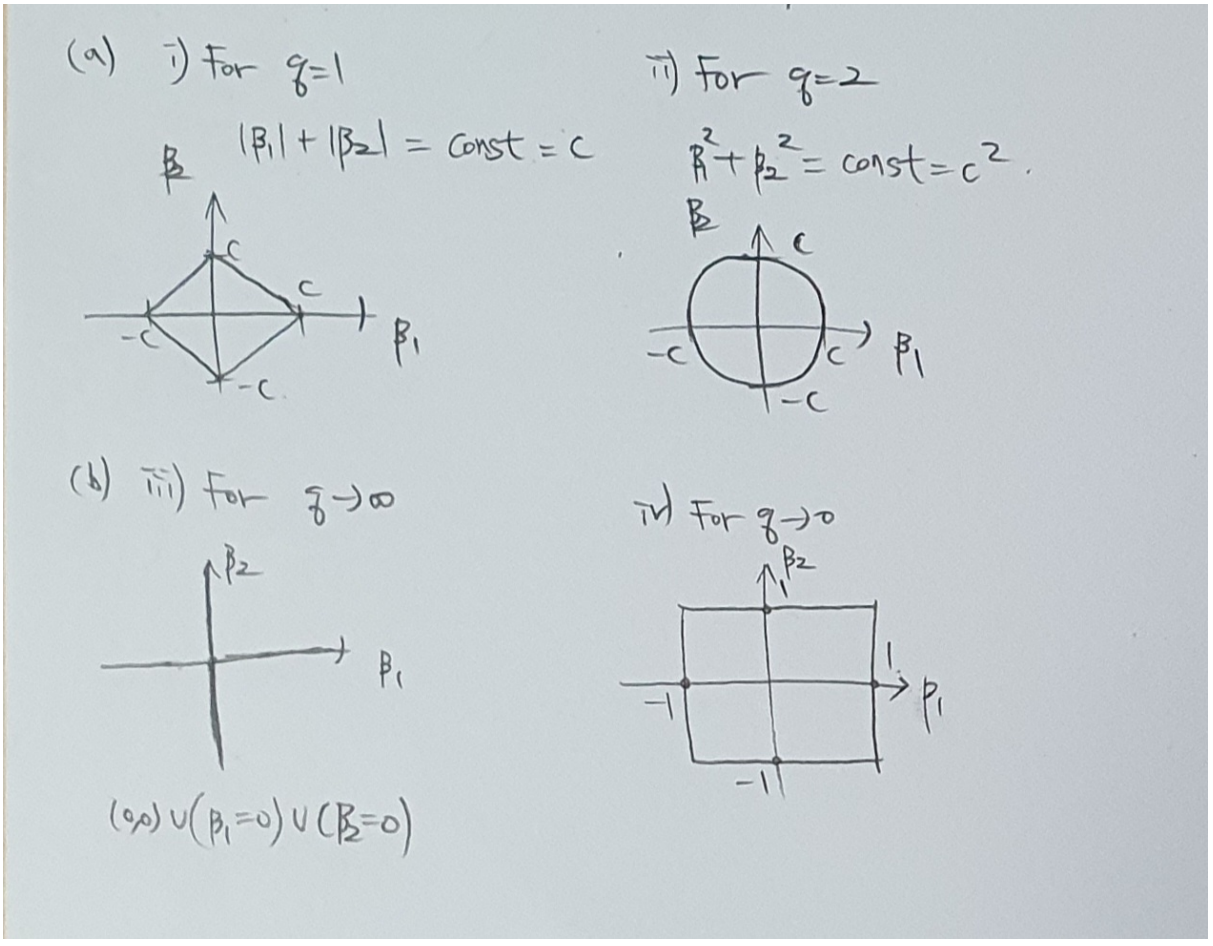
Q5. Drawing contours

In lasso and ridge regression, the term below is known as the Lq norm (or quasi-norm when 0 < q < 1) of the vector β. This term is used in various regularization techniques in statistics and machine learning. When visualized in two dimensions (i.e., considering only β1 and β2), the lines of equal value (contours) for this regularization term assume different shapes based on the q value.

$$\Sigma_j |\beta_j|^q$$

Sketch how these contours transform with varying values of q.
[Note] Do not use any software.
(a) Draw contours when q=1 (lasso) and q=2 (ridge), respectively.

(b) How does the contour change when q → ∞? and q → 0?



Q6. Thoughts on k-fold cross-validation

Consider the challenges of using k-fold cross-validation for time series or temporal data. Why might standard k-fold cross-validation be inappropriate in this context?

(a) How does the temporal ordering of data points in a time series differ from the assumptions made by standard k-fold cross-validation regarding data independence? [3 pts]

Sol) To use standard k-fold-cross-validation, assume that the data is independent by the sequence and identically distributed. However, time series affect before data. That means data is dependent on the sequence. Also, time series have trends and seasonality. That means statistical properties such as the mean and variance, change over time. Thus, data is not identically distributed.

(b) If we were to use standard k-fold cross-validation on a time series dataset, what potential problems might arise during the training and testing phases? [3 pts]

Sol) In k-fold-cross-validation on a time series dataset, training data sets and testing data sets can be mixed. That means the model trains future data. Thus, the model performs well during cross-validation. However, the model fails to unseen future data. In other words, evaluation can be an over-optimistic estimate. Also, k-fold-cross-validation cannot be handled effectively by trends or seasonality.

(c) What modifications or alternative cross-validation techniques could be employed to ensure that the temporal structure of the data is preserved during the validation process? [3 pts]

Sol) To ensure that the temporal structure of the data is preserved during the validation process, use a rolling origin method or expanding window method. For example, there are Walk-Forward Validation, Leave-P-Out Cross-Validation, Block Cross-Validation, and Sliding Window Cross-Validation.

(d) Can you think of a practical scenario in which forecasting future outcomes based on historical data is essential? In your chosen scenario, describe how employing standard k-fold cross-validation could lead to unreliable model predictions. What specific challenges might arise from not respecting the temporal order of the data? [6 pts]

Sol) Consider a scenario that forecasts the financial data, for example, predicting stock prices based on historical data. If using the standard k-fold-cross-validation, due to the over-optimistic estimate in validation, real data prediction could be poor. Also, it does not consider the economic events that affect stock prices. That misestimation leads to potential financial losses.

Q7. Bootstrap

(a) Based on this data set, provide an estimate for the population means of medv data. Call this estimate μˆ. [1 pts]

22.532806324110677

(b) Provide an estimate of the standard error of μˆ. Interpret this result. [2 pts]
Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

0.4084569346972866

(c) Now estimate the standard error of μˆ using the bootstrap. How does this compare to your answer from (b)? [2 pts]

0.42317319229309647

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of medv. Compare it to the results obtained by using Boston['medv'].std() and the two standard error rules. [2 pts]
Hint: You can approximate a 95% confidence interval using the formula [ˆμ − 2SE(ˆμ), μˆ + 2SE(ˆμ)]. Two standard error rule: For linear regression, the 95% confidence interval for β1 approximately takes the form βˆ1 ± 2 · SE(βˆ1).

[21.686459939524482, 23.37915270869687]

(e) Based on this data set, provide an estimate μˆmed, for the median value of medv in the population. [1 pts]

21.2

(f) We now would like to estimate the standard error of μˆmed. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings. [2 pts]

0.3830205085892918

(g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston census tracts. Call this quantity μˆ0.1. (You can use the np.percentile() function. [1 pts]

12.75

(h) Use the bootstrap to estimate the standard error of μˆ0.1. Comment on your findings. [2 pts]

0.5103346720535457