Non-Assessed Exercise

UNIVERSITY OF MANCHESTER
DEPARTMENT OF COMPUTER SCIENCE

DATA70121: Machine Learning and Statistics I

Lecture 10: Regularised Linear Model

Sample Answers

You are strongly suggested making a serious attempt before seeing sample answers.

Lecture 10: Regularised Linear Model

Multiple Choice Questions

1. For any given training dataset, the *ordinary least squares* (OLS) algorithm always produces a unique solution to linear regression. True or False?

   A. True

   B. False

   B

2. A smaller value of the tuning hyperparameter (lambda) in LASSO or Ridge regression leads to more complex models. True or False?

   A. True

   B. False

   A

3. When there are more training examples than input features, the trained Ridge regression model always has a higher RSS than that of the trained OLS model on the training set. True or False?

   A. True

   B. False

   A

4. Both Ridge and LASSO regressions require feature scaling before model training. True or False?

   A. True

   B. False

   A

5. To apply a regularised linear model for regression, input features must be *normalised* to a range [0, 1]. True or False?

   A. True

   B. False

   B

6. The Ridge regression penalty term can be negative. True or False?

   A. True

   B. False

   B

7. What is the effect of the regularisation term in the loss function of a regularised linear regression model?

   A. It penalises a large number of features in the model.

   B. It penalises a small number of features in the model.

   C. It penalises large weights in the model.

   D. It penalises small weights in the model.

   C

8. Which form of regularisation is more likely to be useful when dealing with a dataset that has highly correlated features?

   A. L1 regularisation

   B. L2 regularisation

   C. Both L1 and L2 equally

   D. Neither L1 nor L2

   B

9. If the value of the regularisation parameter (λ) is set to zero, what happens to a regularised linear regression model?

   A. The model becomes fully regularised.

   B. The model becomes equivalent to OLS.

   C. The model's weights are all set to zero.

   D. The model's weights are all set to the biggest weight.

   B

10. In the context of LASSO and Ridge regression, what is the role of the regularisation parameter (λ)?

    A. It determines the learning rate of the optimisation algorithm.

    B. It determines the number of features included in the model.

    C. It determines the number of iterations for the optimisation algorithm.

    D. It determines the complexity of the model.

    D

11. Which of the following is NOT a characteristic of Ridge regression?

    A. Ridge regression shrinks coefficients towards zero.

    B. Ridge regression can handle multicollinearity.

    C. Ridge regression sets some coefficients exactly to zero.

    D. Ridge regression uses L2 penalty.

    C

12. Which type of regression model is known to perform feature selection?

    A. Linear regression

    B. Ridge regression

    C. LASSO regression

    D. Polynomial regression

    C

13. Why is LASSO regression sometimes preferred over Ridge regression?

    A. It can handle multicollinearity.

    B. It is less prone to overfitting.

    C. It is more robust to outliers.

    D. It can perform feature selection.

    D

14. In the geometric interpretation of LASSO and Ridge regression, what does the regularisation parameter (λ) represent?

    A. The radius of the constraint region
    B. The number of dimensions in the feature space
    C. The angle between the coefficients vector and the gradient vector
    D. The distance from the origin to the optimal coefficients vector

    A

15. Which of the following statements best describes the shape of the constraint region in Ridge regression?

    A. It is a hypercube.
    B. It is a hypersphere.
    C. It is a hyperplane.
    D. It is a hypercone.

    B

16. Which of the following statements best describes why LASSO regression can yield sparse solutions?

    A. Because the constraint region is a hypersphere.
    B. Because the constraint region is a hypercube.
    C. Because the constraint region intersects the contour of the loss function at an axis.
    D. Because the constraint region is a hyperplane.

    C

17. What can we infer about the coefficients in LASSO and Ridge regression from their geometric interpretation?

    A. In LASSO, some coefficients can become exactly zero.

    B. In Ridge regression, all coefficients are shrunk towards zero but do not become exactly zero.

    C. In Ridge regression, some coefficients can become exactly zero.

    D. In LASSO, all coefficients are shrunk towards zero but do not become exactly zero.

    A,B

18. Which of the following are reasons to use LASSO or Ridge regression instead of subset selection?

    A. They can perform feature selection and parameter estimation simultaneously because the constraint region is a hypercube.

    B. They are computationally less intensive.

    C. They can handle multicollinearity better.

    D. They can yield interpretable models.

    E. They can handle the case where the number of features is larger than the number of observations.

    A,B,C,D,E

Explanation to Answers

1. *OLS* works only for a dataset that has more training examples (sampled with i.i.d) than the number of features.
2. A smaller λ value means less penalty and thus more complex models, while a larger value leads to a greater penalty and thus simpler models.
3. The *regularisation* term causes a higher RSS than *OLS* on such a training set.
4. Feature with greater numeric ranges could have a larger effect on the penalty term, hence it's a good practice to scale features before training regularised linear regression models.
5. Standardisation does not grantee the range [0, 1].
6. The Ridge regression penalty term is always non-negative. It's the sum of squares of the coefficients, and squares are always non-negative.
7. The purpose of the regularisation term is to penalise large weights in the model, preventing overfitting and encouraging simpler, more general models.
8. Ridge regression (L2 regularisation) tends to distribute weights more evenly among correlated features, making it a better choice for dealing with multicollinearity.
9. If λ is set to zero, the regularisation term in the loss function has no effect, so the model becomes equivalent to OLS without regularisation.
10. The regularisation parameter (λ) determines the strength of the penalty term in the loss function. A larger value results in a greater penalty, which in turn results in a simpler model (i.e., with smaller coefficients).
11. Unlike LASSO regression, Ridge regression does not set any coefficients exactly to zero. It can shrink the coefficients towards zero, but they will not become exactly zero.
12. LASSO regression is known for its ability to perform feature selection, as the L1 penalty can reduce the coefficients of certain features to exactly zero, effectively excluding them from the model.
13. One advantage of LASSO regression over Ridge regression is that it can perform feature selection, which can be helpful when there are many irrelevant features.
14. In the geometric interpretation of LASSO and Ridge regression, the regularisation parameter determines the size or radius of the constraint region. A larger value of λ corresponds to a smaller region.
15. L2 regularisation is used in Ridge regression where the quadratic form of weights leads to the hypersphere shape.
16. The corners of the diamond-shaped constraint region, which represent sparse solutions, are more likely to intersect the contours of the loss function.

17. In LASSO, some coefficients can become exactly zero. In Ridge regression, all coefficients are shrunk towards zero but do not become exactly zero.

18. Five correct answers are explained as follows:

    - **A**: They perform feature selection and parameter estimation simultaneously by shrinking the coefficients of less important features towards zero in Ridge regression and to zero in LASSO.
    - **B**: They are computationally less intensive than evaluating all possible subsets of features.
    - **C**: They can handle multicollinearity better by spreading the effect of correlated features over their coefficients.
    - **D**: They can yield interpretable models, especially lasso, which performs feature selection.
    - **E**: They can handle the case where the number of features is larger than the number of observations, unlike subset selection methods.