Homework 3
Fall 2022 | MDA-710

Due: 10/5 at 11:59pm
Submit on Blackboard

Terminology (40pts)

1. A parameter used as the input to a machine learning model (aka: column names)
2. Previously unseen data used to evaluate the machine learning model
3. Used to convert categorical features to numerical features
4. Used to change the impact of the regularization penalty. Usually denoted by lambda (λ) or alpha (in scikit learn)
5. The sum-of-squares, dynamic penalty used to tackle overfitting
6. A high variance would lead to this
7. Measures the consistency of the model prediction
8. An anonymous function that has no name
9. The expression used in addition to the error function that comprises the cost function
10. The scikit learn function that performs Linear Regression with L1 Regularization
11. The processes of reducing the number of features
12. An algorithm used to assert ML models do not interpret the number representation of nominal features as ordinal
13. Previously unseen data used to adjust the machine learning model
14. The error between average model prediction and the ground truth
15. Data used to create or fit a model
16. A high bias would lead to this
17. When the predictive model is too complex and performs well on the test data, but not the training data
18. The constant penalty used to tackle overfitting by performing feature selection
19. The scikit learn function that performs Linear Regression with L2 Regularization
20. The numerical impact of a feature in a model

A. Feature Selection
B. L1 Regularization
C. L2 Regularization
D. Overfitting
E. Underfitting
F. Bias
G. Variance
H. Penalty
I. Lasso
J. Ridge

K. Lambda function
L. Regularization parameter
M. Feature
N. Weight
O. Test data
P. Validation data
Q. Training Data
R. One-Hot Encoding
S. Mapping

Understanding and extracting useful features

Import the student-exam-score.csv file and write code to perform the following:

1. (10pts) Create mappings for the “parental level of education” feature update your dataframe with the mappings
2. (10pts) Create new training and test data sets from the original data (remember to drop values in the “math-score” column with missing values).
3. (20pts) Using Linear Regression with “math-score” as the dependent variable (y) and “parental level of education” as the independent variable (x), fit the model with your training data and score your model with your test data. Note the score in your answer.
4. (20pts) Using Linear Regression with L1 Regularization, using “math-score” as the dependent variable (y) and “parental level of education” as the independent variable (x), fit the model with each alpha with your training data and score your model with your test data: .1, .15, .3, .6, .9, 1. Note the scores for each alpha in your answer
5. (20pts) Using Linear Regression with L2 Regularization, using “math-score” as the dependent variable (y) and “parental level of education” as the independent variable (x), fit the model with each alpha with your training data and score your model with your test data: .1, .15, .3, .6, .9, 1. Note the scores for each alpha in your answer
6. (30pts) Perform steps 3-5 using the features “math-score” and “reading-score”
7. (30pts) Perform steps 3-5 using the features “math-score” and “reading-score” but divide up you training and test data differently.
8. (5pts) Which feature performed better?
9. (5pts) Which percentage of test data and training data performed better?
10. (10pts) Using the highest performing feature, model (Linear Regression, Linear Regression + L1 Regularization, or Linear Regression + L2) and alpha, update the missing values in ‘math-score` with the predicted values.
11. (10pts) Perform one hot-encoding on the ‘parental level of education’ and add it to the dataframe (remember to drop the original ‘parental level of education’ column).