

Objective: Linear Regression Modeling

Background and problem description

Chemical engineers measured various properties of a gas and created a training data set with four features *temperature (T)*, *pressure (P)*, *thermal conductivity (TC)*, and *sound velocity (SV)*. They also calculated the gas quality using four measured properties and converted it into a quality index (*Idx*). They want to know if any functional relationships exist between the measured properties and the gas quality. The schema of the data set is:


Dataset(T, P, TC, SV, Idx) where the type of each attribute is double

Develop the best predictive model for the training data set using **Python** and write an analysis report.

Required activities

- (1) **(a)** Write a program to learn the best polynomial model using the least square approach by only increasing or decreasing the polynomial order (without using any method to improve the accuracy of a model). **(b)** What's the size of training and testing data in terms of the percentage of the data used? **(c)** What are the training and testing errors in terms of RMSE and R^2 and the training time of the model? **(d)** What's the total number of terms and the polynomial order? **(e)** Show the training results and justify why you think the model is the best.
- (2) **(a)** Write a program **without using any library or package** to learn the best polynomial model using the gradient descent method discussed in class by only increasing or decreasing the polynomial order. **(b)** What's the size of training and testing data in terms of the percentage of the data used? **(c)** What are the initial weights and the learning rate? **(d)** What are the training and testing errors in terms of RMSE and R^2 and the training time of the model? **(e)** What are the total number of terms and the polynomial order? **(f)** Show the training results and justify why you think the model was the best.
- (3) **(a)** Write a program for a feature scaling and run one of the programs written in either (1) or (2) to learn the best polynomial model with the scaled data set. **(b)** What is the feature scaling method used? Show an example row of before and after scaling. **(c)** What are the training and testing errors in terms of RMSE and R^2 and the training time? **(d)** Show the training results with the scaled data and compare the accuracies to determine whether or not scaling impacts the model's accuracy.

(4) **(a)** Write programs to learn the best polynomial model using LASSO, Ridge, and Elastic net discussed in class. **(b)** What are the training and testing errors in terms of RMSE and R^2 for the model learned and the training time by each method. **(c)** What is λ selected for each method with a brief justification of λ ? **(d)** What are the total number of terms and the polynomial order for each model, and what features can be removed and why? **(e)** Select the best model from the three models with a brief justification of the selection. **(f)** Does the regularization help prevent overfitting and ultimately improve accuracy? Justify your answer.



(5) **(a)** Write a program to perform K-fold cross-validation for the three models in (4). **(b)** What value do you choose for K and why? What's the best model selected based on cross-validation? **(c)** Does the cross-validation help select the best model? Justify your answer.

Warning: Although code reuse from source codes available on the Internet is allowed, sharing code with other students (or teams) is strictly prohibited. Any student or team violating this policy will receive a **ZERO** score for this assignment or all the remaining assignments.

What to submit

- Analysis report with your name (or all the members' names), analysis results answering all the questions. If a team completed this assignment and your team does not reach an agreement on individual contribution, briefly write each member's claimed percentage on the specific tasks performed. Different grades for each member may be given based on individual contributions.
- Program file(s)

If the assignment was completed by the team, submit only one for your team.

Grading criteria

- The overall quality of work based on the programs, methods used, results, analysis process, etc.
- The level of understanding shown in the report
- Effort