Applied Machine Learning
Exercise 6

Prof. Dr. Amr Alanwar

November 2024

Datasets

Regression Datasets

1. Generate a Sample dataset called D_1:
   (a) Initialize matrix $x \in \mathbb{R}^{100\times1}$ using normal distribution with $\mu = 1$ and $\sigma = 0.05$.
   (b) Generate target $y \in \mathbb{R}^{100\times1}$ using $y = 1.3x^2 + 4.8x + 8 + \psi$, where $\psi \in \mathbb{R}^{100\times1}$ is randomly initialized.
2. Wine Quality dataset called D_2:
   (use `winequality-red.csv`) Wine Quality Dataset

You are required to pre-process the given datasets.

GLMs: Generalized Linear Models with Scikit-Learn (6 Points)

In previous labs, you have implemented various optimization algorithms to solve linear or logistic regression problems. In this task, you are required to use Scikit-Learn to experiment with the following linear models and Stochastic Gradient Descent (SGD) [Hint: use `SGDRegressor`]:

1. Ordinary Least Squares
2. Ridge Regression
3. LASSO

Following are required in this task:

1. Split your data into Train and Test Splits. Use dataset D_2.
2. For each model, pick three sets of hyperparameters and learn each model (without cross-validation). Measure Train and Test RMSE and plot it on one plot. Explain the plots and relate them to the influence of regularized vs. non-regularized models. You have to compare the models and explain underfitting and overfitting.
3. Tune the hyperparameters using `GridSearchCV` and plot the results of cross-validation for each model. [Hint: use `cv_results_` to see different options].
4. Using the optimal hyperparameter, evaluate each model using `cross_val_score`. Plot each model using a boxplot and explain the significance of your results.

Polynomial Regression (4 Points)

In this task, you are required to use dataset D_1. So far, we have only looked at 1st-degree polynomials (i.e., linear). In this task, you have to use higher degrees of polynomial features for your data: degrees 1, 2, 7, 10, 16, and 100. [Hint: use `sklearn.preprocessing` to generate polynomial features].

Tasks:

1. **Task A: Prediction with High Degree Polynomials**
   (a) For each newly created dataset, learn `LinearRegression`.
   (b) Plot prediction curves for each reprocessed data and (y vs x). Describe the phenomena you observe for different prediction curves.
2. **Task B: Effect of Regularization**
   (a) Fix the degree of the polynomial to 10.
   (b) Pick four values of λ (regularization constant) and learn Ridge Regression [Hint: use `Ridge` and select λ values far apart, e.g., 0, 10^{-6}, 10^{-2}, 1].
   (c) Plot prediction curves for each reprocessed data and (y vs x). Describe the phenomena observed for different prediction curves.

Annex

You can use libraries:

- Scikit-Learn User Guide
- sklearn.metrics Documentation
- sklearn.model_selection
- sklearn.linear_model
- sklearn.preprocessing
- `matplotlib` for plotting.