Lab Course Machine Learning
Exercise 6

Prof. Dr. Dr. Lars Schmidt-Thieme,
Mofassir ul Islam Arif
Information Systems and Machine Learning Lab
University of Hildesheim
Submission: 16.12.2018 LearnWeb 3112

December 9, 2018

1 Exercise Sheet 6

Datasets

- 1. Regression Datasets

  (a) Generate a Sample dataset called D1 :
  i. Initialize matrix $x \in R^{100\times 1}$ using Uniform distribution with $\mu = 1$ and $\sigma = 0.05$
  ii. Generate target $y \in R^{100\times 1}$ using $y = 1.3x^2 + 4.8x + 8 + \psi$, where $\psi \in R^{100\times 1}$ randomly initialized.
  (b) Wine Quality called D2: (use winequality-red.csv)http://archive.ics.uci.edu/ml/datasets/Wine+Quality

You are required to pre-process given datasets.

2 GLMs

Exercise 1: Generalized Linear Models with Scikit Learn (12 Points)

In previous labs you have implemented various optimization algorithms to solve linear or logistic regression problem. In this task you are required to use Scikit Learn to experiment with 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 D2
- 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 it to the theory studied in lectures i.e. influence of regularized vs non-regularized models. You have to compare the following models and argument should explain underfitting and overfitting.
- 3. Now tune the hyperparameters using scikit learn GridSearchCV and plot the results of cross validation for each model. [Hint: use *cv_results* to see different options]
- 4. Using the optimal hyperparameter you have to evaluate each model using cross_val_score. Plot each model using boxplot and explain how significant are your results.

3 Polynomial Regression

In this task you are required to use dataset D1. So far we have only looked at 1st degree polynomial, i.e. linear polynomial and your D1 is also generated using linear polynomial. In this task you have to use more degrees of polynomial feature for your data i.e. degrees 1, 2, 7, 10, 16 and 100. [Hint: use sklearn.preprocessing to generate polynomial features]. Your tasks are:

- 1. **Task A**: Prediction with high degree of polynomials
  - (a) For each newly created dataset learn LinearRegression.
  - (b) Plot prediction curves for each reprocessed data and (y vs x). Which phenomena you observed for different prediction curves.
- 2. **Task B**: Effect of Regularization
  - (a) Fixed the degree of polynomial to 10
  - (b) Pick Four values of λ (regularization constant) and learn Ridge Regression [Hint: use Ridge and your λ values should be far a part i.e. 0, $10^{-6}, 10^{-2}$, 1].
  - (c) Plot prediction curves for each reprocessed data and (y vs x). Which phenomena you observed for different prediction curves.

3.1 ANNEX

- Following lecture is relevant this exercise https://www.ismll.uni-hildesheim.de/lehre/ml-16w/script/ml-04-A3-regularization.pdf
- sklearn.model_selection, sklearn.metrics, sklearn.linear_model, sklearn.preprocessing
- Scikit Learn User Guide http://scikit-learn.org/stable/user_guide.html

- You can use matplotlib for plotting.
- sklearn.metrics http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics