Lab Course Machine Learning
Exercise Sheet 6

Prof. Dr. Dr. Lars Schmidt-Thieme, Hadi Samer Jomaa
Information Systems and Machine Learning Lab
University of Hildesheim

December 4th, 2016
Submission on December 11th, 2016 at 8:00 am, (on moodle, course code 3113)

Instructions

Please read the lab related instructions, i.e. submission, report format and policies, at `https://www.ismll.uni-hildesheim.de/lehre/prakAIML-16w/exercises/ml_lab_instructions.pdf`

Datasets

1. **Regression Datasets**
   (a) Generate a Sample dataset called D_1 :
      i. Initialize matrix $\mathbf{x} \in \mathcal{R}^{100\times 1}$ using Uniform distribution with μ =1 and σ = 0.05
      ii. Generate target $\mathbf{y} \in \mathcal{R}^{100\times 1}$ using $\rightarrow \mathbf{y} = 1.3\mathbf{x}^2 + 4.8\mathbf{x} + 8 + \xi$, where $\xi \in \mathcal{R}^{100\times 1}$ randomly initialized.
   (b) Wine Quality called D_2: (use winequality-red.csv)`http://archive.ics.uci.edu/ml/datasets/Wine+Quality`

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 $\mathcal{D}_2$
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 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.

Exercise 2: Polynomial Regression (8 Points)

In this task you are required to use dataset $\mathcal{D}_1$. So far we have only looked at 1st degree polynomial, i.e. linear polynomial and your $\mathcal{D}_1$ 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.

Bonus: Implement Elastic Net using Stochastic Gradient Descent (SGD)(5 Points)

Elastic Net is a linear model that have both L1 and L2 regularization terms. In this task you are required to implement **(without using Scikit Learn)** Elastic Net model using SGD algorithm. Use dataset $\mathcal{D}_2$.You have to perform

1. Create Train and Test splits.
2. Implement Elastic Net model with SGD.
3. You have to observe the behavior of two regularization constants i.e. λ_1 and λ_2. Choose combination such that 1) Both have small values, 2) λ_1 is zero, 3) λ_2 is zero and 4) larger values. Explain the behavior.
4. Plot learning curves (RMSE on Train and Test for each iteration).
5. Optimize the hyperparameters for this model using Cross Validation.

Annex

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

4. GridSearchCV http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV
5. sklearn.metrics http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics