Machine Learning

Programming Assignment I

Thanks to: Ujjwal Sharma, Shuai Wang, Hongyi Zhu, and Ilker Birbil

The following assignment will test your understanding of topics covered in the first two weeks of the course. This assignment will count towards your grade and should be submitted through Canvas by **19.11.2023 at 23:59 (CET).** You must submit this assignment in teams of 3. You can get at most 10 points for this assignment, which is 10% of your final grade.

Instructions

- Alongside the code for your experiments, you are also required to present a report summarizing the observations and results of each of the experiments. You can use text and graphs/plots (matplotlib) for these reports. You should place these report blocks within the Jupyter Notebook in separate text cells. Plots can be appropriately placed near the text explanation. Your final submission should be a single Jupyter Notebook with code and report blocks.
- While it is perfectly acceptable to brainstorm and discuss solutions with other colleagues, please do not copy code. We will check all submissions for code similarity with each other and with openly-available solutions on the web. Submissions with high similarity will be summarily rejected and no points will be awarded.
- Please ensure that all code blocks are functional before you finalize your submission. Points will NOT be awarded for exercises where code blocks are non-functional.

Submission

You can submit your solutions within a Jupyter Notebook (*.ipynb). To test the code we will use Anaconda Python (3.9). Please state the names and student ids of the authors at the top of the submitted file.

1) Data

In winequality.zip, you will find three files — *winequality.names*, *winequality-red.csv* and *winequality-white.csv*. ‘‘*winequality.names*’’ contains a description of the dataset and the remaining files contain data. These files contain data from a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars [Cortez et al., 2009]. Read the description file carefully. The data consists of 11 *features* containing various sensory and physicochemical properties of wine and a single target variable *"Quality"* (median of at least 3 evaluations made by wine experts). You will find pandas extremely helpful for working with this data. For each experiment, you are also required to split the data into train, validation (if needed) and test splits. Choose an appropriate split ratio.

Before you start fitting models on this data, please analyze the data using Pandas. *DataFrame* methods like *info* and *describe* can provide helpful summaries on the structure and statistics of the data. As a preprocessing step, you are asked to analyze the data and perform the following operations:

1. Convert all features to an appropriate data type. Please read the description and ascertain what these data types should be based on the nature of the information contained in them.
2. Use appropriate plots to demonstrate the distribution of the features and the target variable and check for correlations between different features. In less than 50 words, summarize your findings.

2) Closed-form OLS Solution

Consider a training set comprising N observations of x, written as $X = (x1, \ldots, N)^T$ with corresponding target values $Y = (y_1, \ldots, y_N)$. We assume these data points are individually drawn from the distribution (i.i.d samples). The objective function to be minimized in ordinary linear regression is the sum of squares of residuals:

$$f(\beta) = (Y - \beta^T X)^T (Y - \beta^T X) \qquad (1)$$

In order to find the value of β that minimizes f (β), we differentiate Equation (1) w.r.t β and set it to zero. This yields the following closed-form solution for β:

$$\beta^{OLS} = (X^T X)^{-1} X^T Y \qquad (2)$$

You are asked to perform the following experiments:

1. Generate a random regression problem using the *sklearn.datasets.make_regression* utility. Use a suitable number of features and samples. Crucially, in *make_regression*, set *coef* to *True* to obtain the coefficients of the underlying linear model. Save these coefficients in a variable *coeffs*.
2. Use the optimal parameters obtained via the closed-form solution (eq. (2)) to obtain the set of parameters that minimize the least squares objective.
3. Verify if the parameters obtained from the closed-form solution are the same as *coeffs* and report your results. What is a potential shortcoming of the closed-form approach?

3) Wine Quality Regression

For this assignment, you will implement the following regressors on the wine-quality data:

1. An Ordinary Least-Squares linear regression model.
2. A Ridge regression model that adds L2 regularization to the Ordinary Least-Squares model.
3. A Lasso regression model that adds L1 regularization to the Ordinary Least-Squares model.
4. An ElasticNet regression model with combined L1 and L2 priors as regularizer.

You are asked to perform the following tasks:

1. Fit the *Ordinary Least-Squares* model to the data. Once completed, report the *Mean Squared Error*, *Mean Absolute Error* and the R^2 coefficient of determination.
2. Fit the Ridge, Lasso and ElasticNet regression models to the data. To find an optimal value for the *alpha* and *l1_ratio* (for ElasticNet) hyper-parameters, use the scikit-learn grid search functionality in *sklearn.model_selection.GridSearchCV*. Only the training (and validation; if needed) set should be used for the grid search. You will need to compute the optimal hyperparamaters separately for all models. Report the best *alpha* and *l1_ratio* (where applicable) values from your search. Please provide learning curves and plots to illustrate the effect of the choice of these hyper-parameters on model performance.

Tip:

1. All models required for this assignment can be found in *sklearn.linear_model*.
2. For grid search over hyperparameters, you are advised to consult the *sklearn* documentation to check the default value for that hyperparameter and devise a suitable search strategy.

Grading

Experiment	**Points**
Closed-form OLS	2
OLS Regression	2
Lasso, ridge and ElasticNet Regression	2
Grid Search and Cross Validation	2
Report and Code Quality	2

References

[Cortez et al., 2009] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553.