Assignment 1 consists of two parts – writing and coding. Students can write computer programs for both parts to double-check the solution. However, the writing part should list all algorithm steps with intermediate values for full credit. The same type of problems can be expected on the midterm exam. The programming part can be submitted either as IPython Notebooks (recommended) or as stand-alone scripts. Python interpreter and imported libraries should be compatible with the latest Anaconda distribution (https://www.anaconda.com/).

Writing (30 points total)

Use a precision of two points after decimal (10^{-2}).

1. **(10 point)** Problem 1

Consider the following set of values:

X	Y
-1	0
0	2
1	4
2	5

Provide an optimal linear regression model with intercept using a closed-form matrix formula. Plot it along with the original dataset.

2. **(10 point)** Problem 2

Calculate the Mean Squared Error (MSE) and Mean Absolute Error (MAE) for Problem 1.

3. **(10 point)** Problem 3

Estimate bias and variance of the model from Problem 1.

Programming (70 points total)

1. Download 'Automobile' dataset from the UCI ML repository (https://archive.ics.uci.edu/dataset/10/automobile)
2. Create pandas dataset with the following columns:

Features					Target (Response)
X_1	X_2	X_3	X_4	X_5	Y
wheel-base	compression-ratio	engine-size	length	width	city-mpg

Use the last row (row 3) for the column names.

3. **(5 point)** Designate the first 60% of data as a training set, the next 20% as a validation set, and the last 20% as a test set. You can keep column Y as a part of dataset or separate it into a stand-alone vector.
4. **(10 point)** Consider three linear models for the regression problem – linear regression, ridge regression, and LASSO. Fit them with default parameters on the training set and estimate performance on both validation and test sets. Use MSE, Pearson Correlation Coefficient (PCC), and Coefficient of determination (R^2) metrics.
5. **(20 point)** For ridge regression and LASSO investigate the following values of parameter *alpha* (multiplication coefficient for regularization) on the validation set and its effect on model performance: [0., 0.25, 0.5, 1., 1000.]. Which value gives the best performance on the validation set? Retrain the model with this value and calculate the same metrics on the test set.
6. **(5 point)** Apply Scikit-Learn function `PolynomialFeatures` to the feature part of the dataset (columns X_1- X_5), use the degree 5. Column Y should not be transformed!
7. **(20 point)** Repeat experiments in section 5.
8. **(10 point)** Analyze coefficients of Ridge regression and LASSO models. What is the most important feature (a feature having the largest weight)? Which features have weights close to zero?