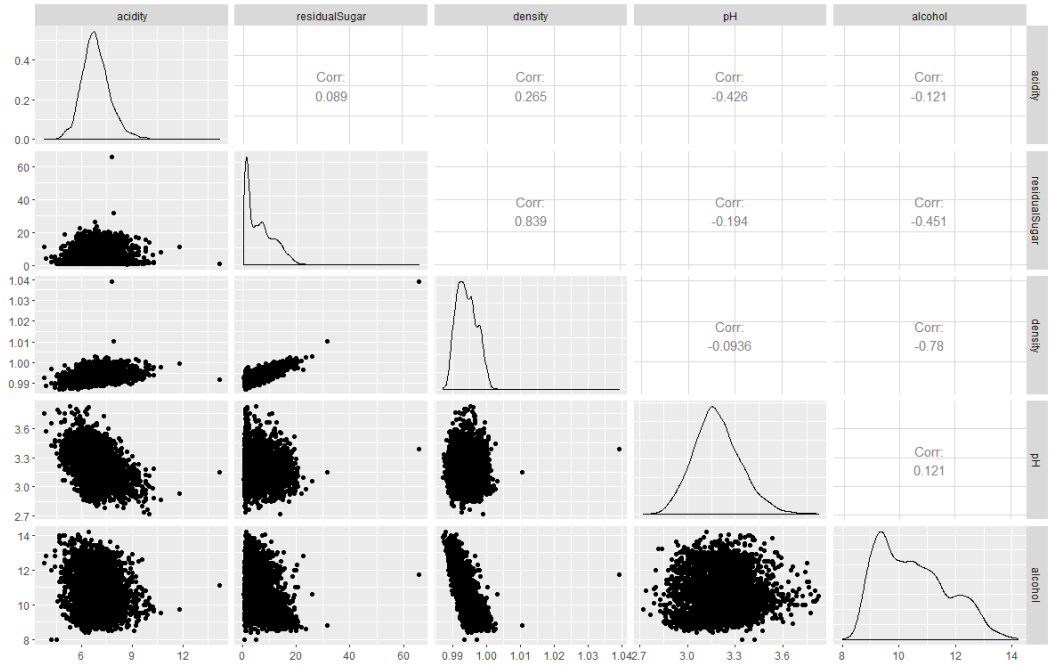
INTRODUCTION TO MACHINE LEARNING

Problem set 2

Burkhardt Funk – Winter 2020/2021

Task 1

Import the CSV file "assignment2wine.csv" (from moodle, modified from the original source: https://onlinecourses.science.psu.edu/stat857/node/223). The dataset contains different characteristics of 4.898 red wines from Portugal.



In this task, you shall explain the *alcohol* of a wine depending on the remaining characteristics (*acidity, residualSugar, density, pH)* assuming a linear dependency. To implement the linear regression use the closed form solution that we derived from the normal equation:

$$\mathrm{w} = \mathrm{X}^{\dagger}\mathrm{y} = (X^T X)^{-1} X^T y$$

Determine the parameter values for *w* ("results are consistent with physics and chemistry"). Compare your result with the LinearRegression class from the sklearn.linear_model module. (The attribute "coef_" contains the estimated coefficients of the model.)

Split the data into a test and train set using the "train_test_split" function from the sklearn.model_selection module using the random_state 99. Train your linear regression model from part a) and write a function that predicts the *residualSugar* on the test set. What is the mean squared error of your predictions? (You can use the mean_squared_error function from the sklearn.metrics module)

Task 2

a) Implement the linear regression using the Widrow-Hoff-Algorithm (i) in batch mode and (ii) for stochastic gradient descent.

b) To apply the algorithm select the attributes *residualSugar* and *density* to explain *alcohol*. Graphically represent the loss function as a 2d contour plot. What do you see? Explore

how normalizing (i.e. “for each variable subtract its mean and divide by its standard deviation”) the input data X changes the contour plot.