MSc in Data Science

Machine Learning

Academic Year: 2017-2018

Exercise 1: Regression and Classification

Delivery Date: **17/11/2017**

You are provided with two datasets, about wine quality. The identity of the dataset can be found in the following link:

- Wine Quality Dataset Identity: http://mlr.cs.umass.edu/ml/machine-learning-databases/wine-quality/winequality.names
- The dataset can be found in the following link: http://mlr.cs.umass.edu/ml/machine-learning-databases/wine-quality/
- Red wine: http://mlr.cs.umass.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv
- White wine: http://mlr.cs.umass.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

Using this dataset, you are requested to learn a set of models according to the following restrictions:

a) **Classification**
Treating the problem as a classification problem, use a decision tree to learn a classification model that predicts wine quality based on the available features. Ensuring that overfitting has not occurred, use the learned model to identify the two most prominent features.

b) **Linear regression**
Treating the learning task as a regression problem, develop a linear regression object that predicts wine quality from all the available features. Perform the experiment 3 times, each time with a different a, and plot the cost function with respect to the training epochs required for the model to converge. Which value of a has been more suitable and why? Does scaling/standardisation affect a? For each one of the prominent features selected in step a), plot the cost function with respect to $\vec{w}$. Finally, describe your processing workflow for modelling the data.

c) **Logistic regression**
Treating the problem as a classification problem, apply logistic regression. For each one of the prominent features selected in step a), plot the cost function with respect to $\vec{w}$, and compare it to least squares cost function with respect to $\vec{w}$. Which of the two cost functions is more suitable for logistic regression and why?