Module Big Data & Small Data
Apply supervised machine learning
Assignment B: Logistic regression

Description of assignment

In logistic regression we use the principle of gradient descent learned in the linear regression assignment and use it to classify data into yes-no predictions rather than predicting a continuous value.

In this assignment you are allowed to change the run script for easier debugging. The signatures of the empty functions "logistic" and "classify" may not be changed, but you are allowed to add new functions to the end of the script.

In this assignment we're looking at a dataset about heart disease; the goal will be to predict the presence of heart disease (1 for heart disease is present, 0 for heart disease is not present) from the measured features. The data set has been split into two parts, one for training, one for validation. The run script loads this data and runs the validation.

Inside the script you will find 3 assignments:

1. Implement logistic regression. At the end of the script you find an empty function named `logistic`. The function takes as input: the feature matrix X, the output vector y, the learning rate alpha and the maximum number of iterations. As output the function should give the computed parameter vector theta and a vector containing the values of the cost function.

2. Experiment and find suitable values for the iterations and the learning rate.

3. Show the classification. At the end of the of the script you find the empty function `classify`. The function takes as input the computed parameter vector theta and the feature matrix X. As output the function should give the classification according to the acquired model.

4. The provided Matlab script uses the accuracy as a metric to evaluate the obtained model. Discuss whether this is a good metric for the given application, heart disease, or would another metric be more useful?

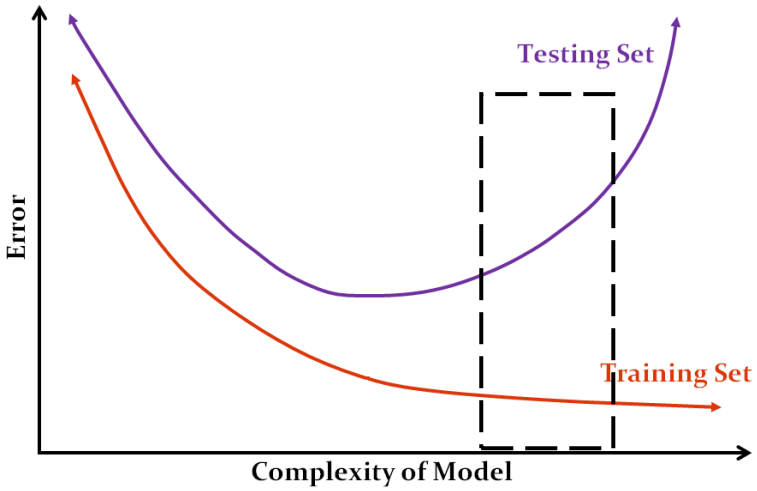

Figure 1 Indication of training and test error as function of the model complexity.

5. Figure 1 gives a sketch of how the error of the training and test set depend on the model complexity. Let's assume the model you have made gives a combination of training and test errors that are situated in the area indicated by the dashed box. Explain what problem your model has in this scenario, and how you can improve your model to overcome this problem.

Files needed:

heart_test.mat
heart_train.mat
run_log_reg.m