Lehrstuhl für Betriebswirtschaftslehre
insbesondere Wirtschaftsinformatik im Dienstleistungsbereich
Univ.-Prof. Dr.-Ing. Wolfgang Maaß

UNIVERSITÄT
DES
SAARLANDES

Voluntary Homework 3

Data Science (Summer term 2020)

Exercise 1 Theory

1. Regularization and Sparsity
   a. A sparse matrix is a matrix where the majority of the values are zero. The proportion of zero elements to non-zero elements is called the sparsity of the matrix. Discuss the relation between regularization (L1 and L2) and sparsity. Use gradients to build your case.
   b. Two regularization techniques namely L1 and L2 have been discussed in the lecture. Explain which among them inherently acts as a feature selector.
2. Decision Trees are a class of powerful classification and regression tree capable of achieving high accuracy while being interpretable. List at least two ways of avoiding / reducing overfitting in Decision Trees. Give reasoning for your choices.
3. Regularization is an effective tool to deal with overfitting. Explain whether input features should be standardized before fitting a regularized model.

Exercise 2 Practical (Regularization and Gradient Descent)

2.1. Read the description of and get familiar with the “housing.csv” dataset. The task is to create a predictive model that can predict the price of a house (last column in the provided dataset) given some of its descriptors.

2.2. Prepare the data of the “housing.csv” dataset for model fitting. Split the data into 60% training, 20% validation and 20% testing parts, and normalize it.

2.3. Complete the code of the class “LinReg” in the file lin_reg.py. Specifically,

* Modify _predict function such that for input matrix X it outputs the numpy vector v, where for every row x_i the value of v_i is $v_i = b + \text{sum } x_{ij} * w_j$, for all j. Values of weights w for weighted sum and bias b are passed to the function together as vector p.

* Modify function “obj” defined in function fit, such that it computes the loss (sum of squared deviations from ground truth) of predictions of the model. Return loss + *C**regularization if *C* is defined and only loss otherwise.

* Use the following 3 regularization scenarios and report *coefficient of determination r2.*

a) No regularization

b) L1 regularization

c) L2 regularization

2.4. Examine what happens to the model weights and to score of the model on test set if you use regularization with value of *C* in [0.01, 1, 100]. Write your observations in exercise_2_4.txt and give explanation to what is happening.