UNIVERSITÄT MANNHEIM

Data and Web Science Group
Prof. Dr. Heiko Paulheim
B6, 26 – B022
68159 Mannheim

Data Mining I

Exercise 6: Regression

6.1. Modeling characteristics of fish

The Fish dataset is a simple dataset which helps to illustrate the linear and non-linear dependencies which may exist between different attributes of the data. The dataset is provided in the *fish.csv* file. It contains 44 examples, each with four attributes: *age*, *water temperature*, *weight* and *length*.

1. Load the Fish dataset and visualize it by combining different attributes. Can you make an assumption about the function to predict the length and weight based on one of the variables?
2. Learn a linear regression model based on the combination of attributes you find most convenient for length and for weight (one regression using length as label, and one regression using weight as label). Which types of regression work best? Do they apply equally to all combinations of attributes?
3. Measure the performance of the different regression models you learned before. Use 10-fold cross validation and RMSE as well as R^2 for evaluation.

6.2. Feature Selection

In this exercise you will explore different feature selection methods for linear regression.

1. First, fit a linear regression model to the “birthweight_train” dataset without any feature selection and evaluate the model on the “birthweight_test” dataset.
2. Look at the resulting model and inspect the p-values for each feature. Fit a second regression model using only the significant features ($p<=0.05$). How does the performance of your model change?
3. Look at the new model and inspect the p-values again. Are there any features for which the p-value has changed? Remove them and fit a new regression. Does this improve your performance?
4. Now, use all features again and let the regression figure out which features to use. What you did in subtask 2 corresponds to feature selection with “T-Test”, and subtask 3 corresponds to “Iterative T-Test”. Test both settings and compare the results to your manual feature selection.

6.3. Predicting housing prices in Boston

The Housing dataset describes 506 houses in the suburbs of Boston in 1993. The data set is provided in the *housing.csv* file. The houses are described by the following 12 continuous attributes and 1 binary attribute:

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

1. Your task is to find a good regression model for determining the median value of a house (MEDV). You may experiment with different regression methods and parameter tuning. As always, it may help to first visualize different attribute combinations of the data. What are the best RMSE and R^2 that you can achieve using the different methods?