

Data Mining – FSS 2020

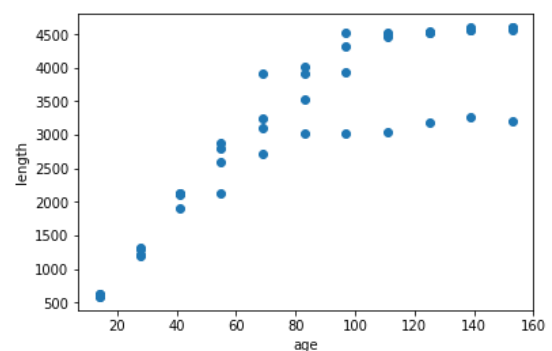
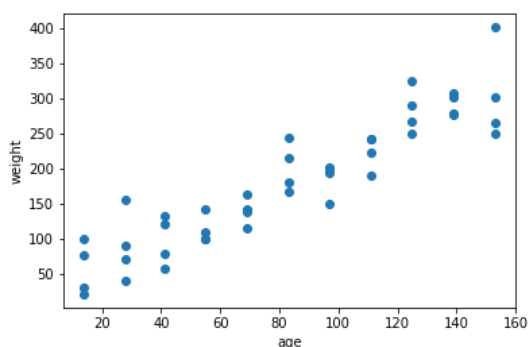
Exercise 8: Regression

8.1. 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*.

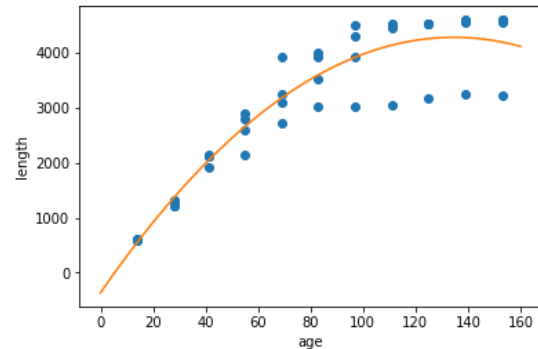
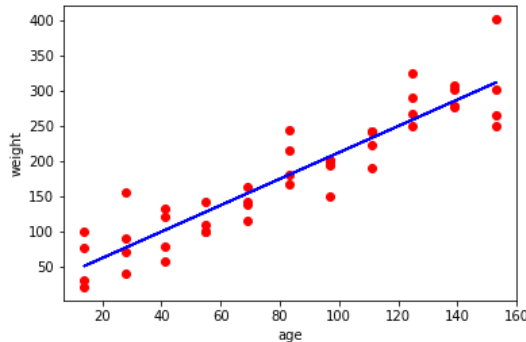
1. Load the Fish dataset and visualize it by combining different attributes. Can you make an assumption about the function to predict the length and weight based on one of the variables?

Solution: When plotting all different attribute combinations (6 in total), we see a very clear linear dependency between the attributes *age* and *weight*. There is also a non-linear, but close to linear, dependency between the attributes *age* and *length*.



- Learn a 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?

Solution:



Prediction “weight”: $24.463 + 1.874age$
MSE: 1266.673 / RMSE: 33.789 / R^2 : 0.798

Predicting “length”: $-393.533 - 0.281age^2 + 70.736age$
MSE: 241052.473 / RMSE: 445.196 / R^2 : 0.751

- Measure the performance of the different regression models you learned before. Use MSE, RMSE, and R^2 for evaluation.**

Solution: The linear model learned for the weight as a function of age has a MSE of 1266.673 and a RMSE of 33.789. The non-linear model learned for the length as a function of age has an MSE of 241052.473 and an RMSE of 445.196. By comparison, if you learn a linear model for the length as a function of the age, you get a RMSE of 631.122, which shows the non-linear models fits the data better.

8.2. Feature Selection

In this exercise you will explore different feature selection methods for linear regression. When importing the data, set the role of the “Birthweight” attribute to “label” and the role of the “id” attribute to “id”.

- First, fit a linear regression model to the “birthweight_train” dataset without any feature selection and evaluate the model on the “birthweight_test” dataset.

Solution: Set up a process that learns a linear regression on the training dataset and evaluate the performance on the test dataset. The RMSE should be 1.060.

- Look at the resulting model and inspect the p-values for each feature. Fit a second regression model using only the significant features ($p \leq 0.05$). How does the performance of your model change?

Solution: Add a “Select Attributes” operator before the regression learner and select the attributes “Birthweight, Gestation, fage, headcircumference, mppwt”. The RMSE should decrease to 0.883. In the new model, only “headcircumference” and “Gestation” are significant features.

- 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?

Solution: Add a “Select Attributes” operator before the regression learner and select the attributes “Birthweight, Gestation, headcircumference”. The RMSE should decrease to 0.725. In the new model, only “headcircumference” and “Gestation” are significant features.

- 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.

8.3. 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

- Your task is to find a good regression model for determining the median value of a house. You may experiment with different regression models and parameter tuning. As always, it may help to first visualize different attribute combinations of the data.

Solution: The following table shows the results which are achievable with different models and configurations:

Model	Configuration	RMSE	R ²
Linear Regression	Default	4.545	0.728
Neural Net	1 layer (5 nodes)	4.140	0.814
Neural Net	1 layer (5 nodes); learning rate=0.1	2.957	0.887
Decision Tree	Default	4.715	0.720
Decision Tree	Depth=7; prepruning=false	4.347	0.757

