Orange Exercises – Part 1

Question 1.

Using Orange, create your own dataset consisting of at least 500 data points (instances). The dataset should contain 3 variables: two of them continuous and one being a categorical variable with 3 discrete values (a class). The categorical variable should be more or less equally distributed (~ 150 instances for each class).

1.1. *Figure presentation:* Present a scatterplot displaying the relationship between the two continuous variables and including the categorical variable.

1.2. *Fitting a model:* Fit a linear model on your data. Describe what you observe: How successful is the fit?

1.3. *Comparing predictors:* Which of the two continuous variables, if any, is a better predictor of the categorical variable? Support your argument with a figure using one of the visualization widgets in Orange.

Question 2.

Using Orange, create two datasets, one illustrating homoscedasticity and one illustrating heteroscedasticity. The datasets should consist of at least 500 data points each and should not represent a simple linear relationship (think, e.g., of a quadratic relation).

2.1. *Figure presentation:* For both datasets separately, provide a figure that illustrates the homoscedasticity and the heteroscedasticity property, respectively (i.e., in total, you will be providing two figures).

2.2. *Fitting a model:* Fit a curve on your dataset that *satisfies* homoscedasticity. Describe what you observe: How successful is the fit? Which value of polynomial expansion did you use and why? Use both text and visual representations to support your description.

2.3. *Sampling*: Conduct data sampling on your dataset *violating* homoscedasticity. Use different sample sizes, one of 5% and the other one of 70%. Compare the distributions and fit a linear model on both of them. What do you observe with respect to the fit? Use both text and visual representations to support your observation.

Question 3.

Create a dataset with a continuous variable that can easily be discretized (binned) into three values. Explain your reasoning and provide a figure displaying the discretization (binning).

Question 4.

Using Orange, explain the concept of *overfitting* in your own words to a data science novice. Your explanation should contain both figures and text (ca. 150-200 words).