Orange Exercises – Part 2

Question 1.

In Orange, load the dataset *zoo.tab* using the File widget. The dataset should contain 17 discrete variables, verify this using the Data info widget.

1.1. *Data exploration*: Answer the following questions about the dataset:

a) Is this a regression or a classification task?

b) How many rows are there in the dataset?

c) Connect the Distributions widget to the File widget, and inspect the distribution for the *Type* feature. What is the most common type of animal?

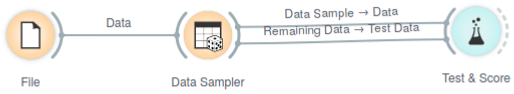
1.2. *Splitting the data:* Connect the File widget to a Data Sampler widget. Choose to sample a Fixed proportion of the data (of 70%), make sure the option for '*Stratify Sample (when possible)*' is off and 'Replicable (deterministic) sampling' is on. Answer the following questions:

a) How many instances are there in the 'Data Sample', and how many in the 'Remaining Data'?

b) For both the 'Data Sample' and the 'Remaining Data', what is the most common type of animal, and what is the second most common type of animal?

c) Now enable the option 'Stratify Sample' in the Data Sampler widget and answer the previous question again (1.2b). Did the distributions of types of animals change? Explain in your own words what happened and whether it is better to have 'Stratify Sample' on or off.

1.3. *Preparing for evaluation*: Add a Test & Score widget and connect the 'Data Sample' and the 'Remaining Data' to it, make sure the 'Data Sample' goes to 'Data', and the 'Remaining Data' to 'Test Data', as in the image below.



Soon we will connect the classifiers to our Test & Score widget so we can start fitting models. But first we need to decide what Sampling option (in the Test & Score widget) to use for fitting our models: should we do cross validation, leave one out, test on train data, or test on test data? Justify your choice for the Sampling option, if you choose cross validation explain how many folds you will use.

Question 2.

Using the workflow we created for question 1 do the following:

2.1. *Adding the models:* add the Logistic Regression, Majority, and Naive Bayes widgets to your workflow and connect them to the Test & Score widget. Make sure the Regularization type for Logistic Regression is set to 'Ridge (L2)', and the Strength such that C=1. Take a first look at the Evaluation Results in the Test & Score widget, how did your models do? Explain which metric you looked at and why you chose that metric.

2.2. *Choosing model parameters:* Which model parameters can we change? If we change any of the parameters should we then look at the training, validation, or test performance to see if this was the right change (keep in mind that we might change the parameters again after this)?

2.3. *Parameter search*: change the Strength/C of the Logistic Regression, what happens to the performance (i.e., the evaluation results in Test & Score)? Try to find the best C value, describe the process of how you tried to find it.

2.4 *Model comparison*: using the sampling option you have used for Test & Score until now, what is the best performing model? Justify your answer using the metric you chose for question 2.1.

2.5 *Unseen data*: evaluate how your models perform on unseen data. Answer the following questions:

a) What is the best performing model on the unseen data? Justify your answer using the metric you chose for question 2.1.

b) Is the model you picked for question 2.4 still the best model? Explain why it is or is not the case, in your own words.

Question 3.

For this task you are required to make a new workflow that mimics the one you made previously, but for a dataset of your choosing. Orange comes with a collection of datasets which you can select from inside Orange, or download here: https://github.com/biolab/orange3/tree/master/Orange/datasets. From these datasets pick a dataset that is intended for classification and has not been used in the course so far (so not: breast-cancer-wisconsin.tab, zoo.tab, iris.tab, or housing.tab).

3.1. *Describe your dataset*: Which dataset did you choose? How many rows does it have? How many discrete and/or continuous features?

3.2. *Data preparation*: split your data into a train and test set using the Data Sampler (with replicable sampling on), connect both sets to the Test & Score widget, and choose a cross-validation strategy. Describe the steps you have taken, and justify your choices based on the characteristics of your dataset.

3.3. *Establish a baseline*: add the appropriate baseline widget to your workflow and connect it to your Test & Score widget. Answer the following questions:

a) Which widget did you choose?

b) Pick a metric you will look at to evaluate the results. What is the baseline performance on this metric?

c) Do you think classification on this dataset is a difficult task or an easy task? Why?

3.4. *Perform classification*: add the Logistic Regression and Naive Bayes widgets to your workflow, and connect them to the Test & Score widget. Answer the following questions:

a) What are the evaluation results for Logistic Regression and Naive Bayes on the metric you chose for question 3.3?

b) Did your classifiers perform better than the baseline?

c) Do you think it is useful to use a classifier for this task? Explain in your own words.