Applied Machine Learning Homework 2

Due 2/20/19 1pm.

Please create the submission using github classroom with the following link:
https://classroom.github.com/a/hxegGwkW

Please submit Task 1 and Task2 as separate Jupyter notebooks. Clearly mark which part of the notebook fulfils which task.
You need to also submit a single PDF containing the content of both notebooks to gradescope.

Task 1 Regression on Ames Housing Dataset (60 points)

You can load the Ames housing dataset from
http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.xls
You can find a description of the variables here:
http://jse.amstat.org/v19n3/decock/DataDocumentation.txt
Take note that for categorical variables, NA here does not mean a missing value, but should be treated as a separate category.

1.1 Visualize the univariate distribution of each continuous, and the distribution of the target. Do you notice anything? Is something that you think might require special treatment (comment what it is, you're not required to try to fix it).

1.2 Visualize the dependency of the target on each continuous feature (2d scatter plot).

1.3 Split data in training and test set. Do not use the test-set unless for a final evaluation in 1.6. For each categorical variable, cross-validate a Linear Regression model using just this variable (one-hot-encoded). Visualize the three categorical variables that provide the best R^2 value.

1.4 Use ColumnTransformer and pipeline to encode categorical variables. Evaluate Linear Regression (OLS), Ridge, Lasso and ElasticNet using cross-validation with the default parameters. Does scaling the data (within the pipeline) with StandardScaler help?

1.5 Tune the parameters of the models using GridSearchCV. Do the results improve? Visualize the dependence of the validation score on the parameters for Ridge, Lasso and ElasticNet.

1.6 Visualize the coefficients of the resulting models. Do they agree on which features are important?

Task 2 Classification on the Telco-churn dataset (40 points)

You can download the dataset and see it’s description at
https://www.kaggle.com/blastchar/telco-customer-churn

2.1 Visualize the univariate distribution of each continuous feature, and the distribution of the target.

2.2 Split data into training and test set. Build a pipeline for dealing with categorical variables. Evaluate Logistic Regression, linear support vector machines and nearest centroids using cross-validation. How different are the results? How does scaling the continuous features with StandardScaler influence the results?

2.3 Tune the parameters using GridSearchCV. Do the results improve?
Visualize the performance as function of the parameters for all three models.

2.4 Change the cross-validation strategy from ‘stratified k-fold’ to ‘kfold’ with shuffling. Do the parameters that are found change? Do they change if you change the random seed of the shuffling? Or if you change the random state of the split into training and test data?

2.5 Visualize the coefficients for LogisticRegression and Linear Support Vector Machines using hyper-parameters that performed well in the grid-search.