

Applied Machine Learning Homework 2

Due 2/17/18 1pm.

Please create the submission using github classroom with the following link:

<https://classroom.github.com/a/HvfDpwbN>

Please submit Task 1 and Task2 as separate Jupyter notebooks. Clearly mark which part of the notebook fulfils which task.

Task 1 Regression on California House prices

You can load the California housing dataset using `sklearn.datasets.fetch_california_housing`.

You can find a description of the data here:

http://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

1.1 Visualize the univariate distribution of each feature, 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 feature (2d scatter plot).

1.3 Split data in training and test set. Evaluate Linear Regression (OLS), Ridge, Lasso and ElasticNet using cross-validation with the default parameters. Make sure to preprocess the data appropriately.

1.4 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.

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

Task 2 Classification on the Covertypes Dataset

You can load the Covertypes dataset using `sklearn.datasets.fetch_covtype`. You can find details about the dataset here: <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

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

2.2 Split data into training and test set. Evaluate Logistic Regression, linear support vector machines and nearest centroids using cross-validation. How different are 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 split into training and test data?

2.5 Visualize the coefficients for LogisticRegression and Linear Support Vector Machines.