Ph464 - Scientific Computing II - Fall 2018

Project 2, due Friday, October 26

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`

a) Visualize the univariate distribution (one random variable) 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, youre not required to try to fix it).

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

c) Split data in training and test set. Evaluate Linear Regression (OLS), Ridge, Lasso and ElasticNet using cross-validation with the default parameters. Does scaling the data with StandardScaler help?.

d) 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.

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

2. Classification on the Covertype Dataset

You can load the Covertype dataset using sklearn.datasets.fetch_covtype. You can find details about the dataset here: `https://archive.ics.uci.edu/ml/datasets/covertype`

a) Visualize the univariate distribution of each feature, and the distribution of the target.

b) 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? How does scaling the data with StandardScaler influence the results?

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

d) 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?

e) Visualize the coefficients for LogisticRegression and Linear Support Vector Machines.