TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Feature Reduction

Programming for Data Science

Evaluation of Predictors

Dresden Database Systems Group

Problem: Overfitting

- Predictor is optimized on training set (fine granularity of rules)
- Bad results on whole dataset/unknown observations
  - Quality of training set (noise, missing value, wrong values)
  - Different statistic properties of training and test set

Problem: Underfitting

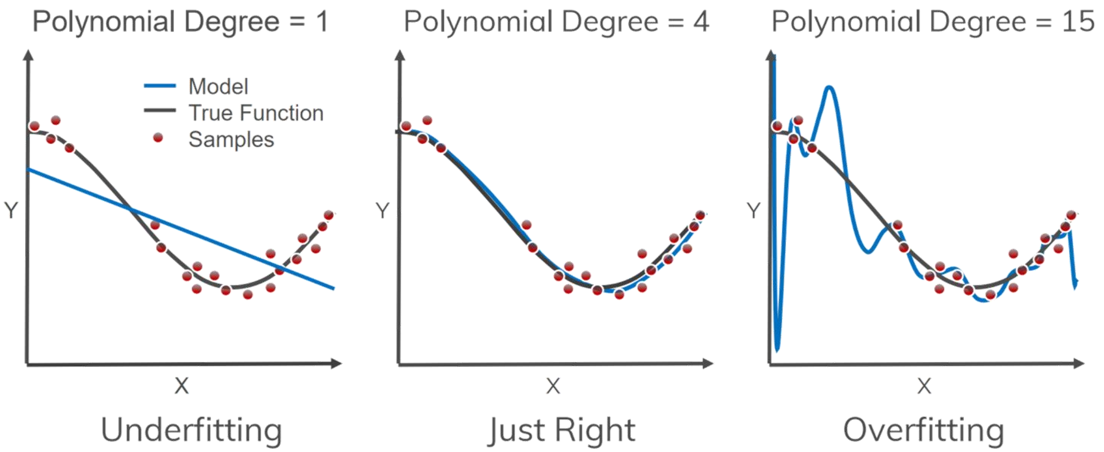
- Predictor is not optimized on training set (coarse granularity of rules)
- Bad results on whole dataset/unknown observations
  - Quality of predictor (dimensionality, parametrization)

Solution: Train-and-Test

- Split data set X into two partitions
  - Training set: Classifier learns with these observations
  - Test set: Evaluation: Prediction of class labels and comparison with existing class labels
- Problem: Not applicable if training set is too small

Evaluation of Predictors (2)

Overfitting vs Underfitting



Occam's razor

- "Plurality must never be posited without necessity"
- A more general solution is always a better solution than a highly adapted solution

Correlation-based Feature Selection

Dresden Database Systems Group

Correlated features do not add information.

- Given a regression $y = a_1x_1 + a_2x_2$ where x_1 and x_2 are highly correlated.

$$\exists\gamma : x_2 \sim \gamma x_1$$

$$y = a_1x_1 + a_2(\gamma x_1)$$
$$y = (a_1 + a_2\gamma)x_1$$
$$y = a_3x_1$$

- Eliminate one of the correlated features.
- Highly correlated usually means $|correlation| > 0.6$

1. Find all (absolute) correlations between features.
2. Select the feature pairs with higher correlation than threshold.
3. Eliminate one of the features (Overall order is important. $a \sim b,\ b \sim c$ should remove b).

Principal Component Analysis (PCA)

Transformation of correlated features to uncorrelated.

- Also used for dimensionality reduction.
- Based on eigenvectors

Target: $\mathrm{T} = XW$, find W

Columns of W are the eigenvectors of $X^T X$

If W is column-truncated $T_p = XW_p$,
T_p is a reduced feature space (p-dimensional)

Task

Step 0

- You will get a csv file from us. Load it in your language/environment.
- Explore the data in it. Identify features and labels.

Step 1

- Split your data into training and test set. Use a ratio of 80-20.
- Use OLS to predict the labels. Report the MSE for training and test set.

Step 2

- Use Lasso regression to automatically reduce the feature space ($\lambda = 0.1$). Report the MSE for the training and test set.

Step 3

- Implement a correlation-based feature selection*. Use 0.6 as a threshold.
- Train OLS on the reduced feature space. Report the MSE for the training and test set.

Step 4

- Implement PCA*. Transform your feature space to 2D.
- Train OLS on the reduced feature space. Report the MSE for the training and test set.

*use your own implementation

Package suggestions

R

- (data.table)
- glmnet
- stats

python3

- numpy
- pandas
- sklearn
- (matplotlib)

Exercise Appointment

Dresden Database Systems Group

We compare and discuss the results

- Tuesday, 05.01.2021,
- Consultation: Please use the forum in Opal.
- Please prepare your solutions! Send us your code!

If you have questions, please mail us:

claudio.hartmann@tu-dresden.de Orga + Code + R
lucas.woltmann@tu-dresden.de Tasks + Python