Prof. Dr. Michael Griebel
Prof. Dr. Jochen Garcke
Dr. Bastian Bohn
Jannik Schürg

6 MACHINE LEARNING PROJECTS

Prepare your presentation until July 18th.

By now you have encountered the most important basic methodologies and techniques needed to analyze given labeled and unlabeled data sets: from data acquisition, preparation and preprocessing to applying state-of-the-art machine learning algorithms for dimensionality reduction, clustering/visualization and classification/regression. For this final sheet/project, you will work freely on a real-world data set to explore and analyze it. Note that you may employ any python machine learning library you want but you should be able to explain and motivate every step you take and provide references if needed.

SUITABLE DATA SETS

Task 6.1. *Pick one of the data sets below and analyze it based on what you have learned during the practical lab course.*

In the following, we present a few possible data sets and ideas you can follow and base your project on.

CIFAR-10 image data set

The CIFAR-10 dataset (downloadable at https://www.cs.toronto.edu/~kriz/cifar.html) consists of 60000 RGB-images of size 32 × 32. Each image is labeled according to what it shows: An airplane, an automobile, a bird, a cat, a deer, a dog, a frog, a horse, a ship or a truck.[1]

By now, you should have enough knowledge in training neural networks, building suitable features and employing dimension reduction methods to do some meaningful analysis of the dataset.

Rhine level data set

The rhine level data set can be downloaded at the website of the practical lab. Here, the water levels of the river rhine and tributaries from 1981 until 2013 are given in 15-minute intervals at 15 different locations. For the analysis of time series data like this it makes sense to look into

1 If you find it boring by now to work with only 10 classes, have a look at the CIFAR-100 dataset on the same website.

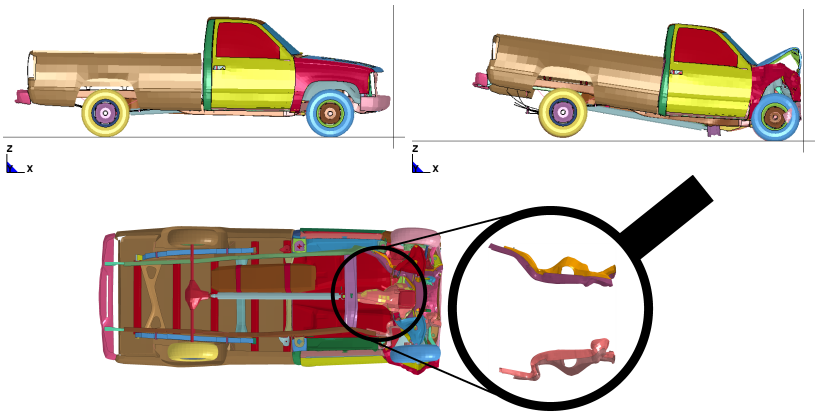

Figure 6.1: Top: The truck finite element model before and after a crash. Bottom: A view from below with the extracted longitudinal beams.

the concepts of *delay embedding* and *(windowed) Fourier features*. A possible task is to predict the water level in Bonn 12 hours ahead by using all past data.

Car crash data set

The car crash data set can be downloaded at the website of the practical lab. Here, 126 frontal car crash simulations of a finite element model of a Chevrolet pickup truck have been computed. The finite element model for every simulation was the same, but certain parameters like the impact velocity or the thickness of some car parts were altered. From the car model we extracted 2 longitudinal beams, which are relevant for the safety behaviour. The given data set contains the file `origPositions.csv` representing the initial positions of each finite element node of the model. Here, each line consists of a number identifying the node and the x-,y- and z-position of the node. Analogously the files `DisplacementSimulationN.csv` contains the displacements of the x-, y- and z-positions after crash N, i.e. when adding the values of the original positions to the DisplacementSimulationN values, you will get the positions of the nodes after crash N.

One possibility to analyze this data set would be to put the values from each displacement file into one large vector such that one has 126 very high-dimensional data points, where each data point corresponds to one crash simulation. Then a dimensionality reduction and/or clustering might reveal some structure of the data set.[2] However, this is only one possibility to approach this data set.

2 Hint: To visualize the beams you can simply use a $3d$ scatter plot for the x-,y- and z-coordinates from origPositions.

20 newsgroups data set

The 20 newsgroups data set is a frequently used benchmark data set for text classification. Here, 11314 training and 7532 test newsgroup texts from 20 different newsgroup categories are to be analyzed. The data set can be downloaded and preprocessed with scikit-learn for instance, see http://scikit-learn.org/stable/datasets/twenty_newsgroups.html. Make sure to strip any revealing metadata before working with it, i.e. use

```
newsgroups_train = fetch_20newsgroups(subset='train',
    remove=('headers', 'footers', 'quotes'))
```

to obtain the data. Besides *tf* and *idf* features as used in the scikit-learn examples, you might want to have a look at more general *bag of words* or *n-gram* features as well as *stemming/lemmatization* and *stop-word-filtering*.

A different data set

You can find many interesting data sets at e.g.

- https://archive.ics.uci.edu/ml/index.php: Several benchmark data sets as they are often used in scientific papers,
- http://www.kaggle.com: Many real world data sets and challenging data mining competitions.

If you already have your own data set in mind that you want to analyze, that's fine too. However, if you do not want to use one of the data sets explained in the sections above, you should briefly consult with us before you start working on your chosen data set to ensure that your project is realizable and meaningful.

FINAL PRESENTATION

You will present the results of your project work in a short presentation at the end of the practical lab course. Furthermore, we will also assign a specific topic of the course to you, which you should also include in the presentation. Make sure that everyone in your working group talks for (approximately) the same amount of time during the presentation.

Task 6.2. *Prepare a 10 minute presentation (approximately 5-10 PDF slides) about your assigned topic and your project work: What did you do? What was your motivation? What results did you get? How would you interprete/compare/rate these results? You should also be able to present and run parts of the code you wrote during the semester upon request. You will give the talk on July, 18th, during the practical lab course. After your talk there will be a brief discussion on your project work and the contents of the practical lab course.*

Be creative and have fun!