Project 2 on Machine Learning, deadline February 20, 2020

Data Analysis and Machine Learning

Jan 30, 2020

Paths for project 2

For project 2, you can propose own data sets that relate to your research interests or just use existing data sets from say

1. Kaggle
2. The University of California at Irvine (UCI) with its machine learning repository
3. The credit card data set from UCI is also interesting and links to a recent scientific article. See however below for possible project example
4. The pulsar classification data set is obtained from Kaggle, where it was posted by Pavan Raj. The data file is available in the DataFiles folder of this project.
5. Or other data sets you find interesting and relevant.

The approach to the analysis of these new data sets should follow to a large extent what you did in project 1. That is: Whether you end up with a regression or a classification problem, you should employ at least two of the methods we have discussed among linear regression (including Ridge and Lasso), Logistic Regression, Neural Networks, Support Vector Machines (not covered during the lectures) and Decision Trees, Random Forests, Bagging and Boosting. If you wish to venture into convolutional neural networks or recurrent neural networks, or extensions of neural networks, feel free to do so. For project 2, you should feel free to write your own code or use the available functionality of scikit-learn, tensorflow, etc.

The estimates you used and tested in project 1 should also be included, that is the R2-score, MSE, cross-validation and/or bootstrap if these are relevant. If possible, you should link the data sets with exisiting research and analyses

thereof. Scientific articles which have used Machine Learning algorithms to analyze the data are highly welcome. Perhaps you can improve previous analyses and even publish a new article? A critical assessment of the methods with ditto perspectives and recommendations is also something you need to include. All in all, the report should follow the same pattern with abstract, introduction, methods, code, results, conclusions etc as in project 1.

Studying the credit card data set as possible project. We include this data set as an example on how one could study new data sets with the algorithms we have discussed during the lectures, using either your own codes or the functionality of scikit-learn, tensorflow or other Python packages.

The data set is presented at the site of UCI. It is particularly interesting since it is also analyzed using ML methods in a recent scientific article.

The authors apply several ML methods, from nearest neighbors via logistic regression to neural networks and Bayesian analysis (not covered much in our course). Here follows a set up on how to analyze these data.

Part a). The first part deals with structuring and reading the data, much along the same lines as done in projects 1 and 2.

Part b). Perform a logistic regression analysis and see if you can reproduce the results of figure 3 of the above article.

Part c). The next step is to use neural networks and the functionality provided by tensorflow/keras or scikit-learn's MLP method (or you could write your own code). Compare and discuss again your results with those from the above article.

Part d). The above article does not study random forests or support vector machine algorithms. Try to apply one of these methods or both to the credit card data and see if these methods provide a better description of the data. Can you outperform the authors of the article?

Part e). Finally, here you should present a critical assessment of the methods you have studied and link your results with the existing literature.

The Pulsar data. The pulsar classification data set is obtained from Kaggle, where it was posted by Pavan Raj. It offers an interesting possible classification problem. In the field of radio astronomy, pulsars are among the most studied phenomena in nature. But despite astronomers' long history with pulsars, little is actually known with certainty. However, much of the uncertainty likely boils down to the difficulty of confirming pulsar observations. While pulsars radiate unmistakable radio signals, they are often lost in the sheer number of radio signals observed by radio telescopes every day. Furthermore, due to the uniqueness of pulsar radio signals, classifying pulsars in large data sets of radio observations

have historically been very difficult as human supervision has been a necessity. However, recent advances in machine learning and data mining has made this task much simpler by introducing incredibly fast, in comparison to humans that is, classification methods.

You could repeat many of the steps discussed for the credit card data problem. The article of Bathes et. al. can serve as a reference for your discussions.