



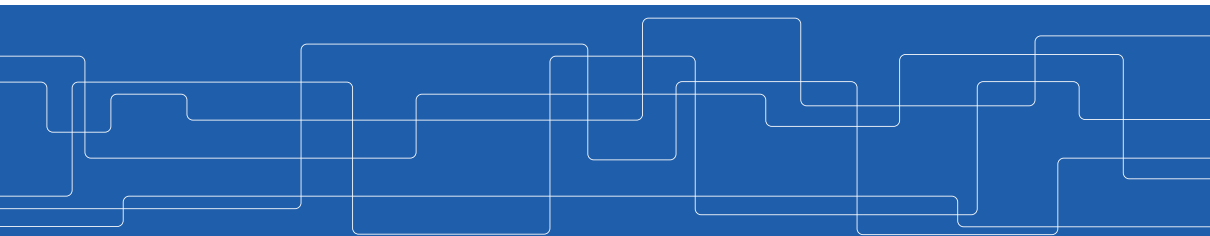
Practical Session

ID2214 Programming for Data Science

Amir A. Rahn timer

`arahnama@kth.se`

`https://gits-15.sys.kth.se/amiakh/ID2214`





Agenda

- ▶ In this session, we will look at a real-world case and try to walk through different stages of a Machine Learning project in a principled way
- ▶ First, we will start by looking at the data and the task
- ▶ After that, we will go through Exploratory data analysis phase
- ▶ In the end, we will show you some ideas about how to go about modelling data you see



Usecase: Netflix Movies and TV Shows

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine. In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled.



Some notes about the dataset

- ▶ You can access the data at <https://www.kaggle.com/shivamb/netflix-shows>
- ▶ It has both text and numerical features
- ▶ Data already is not yet divided into training and test
- ▶ The number of instances 6234 with 12 features

Exploratory Data Analysis (EDA)

- ▶ In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- ▶ Primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was suggested by John Tukey as an approach to encourage statisticians to explore the data, before formulating hypotheses
- ▶ DA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA. Wikipedia [2019]



Exploratory Data Analysis (EDA) Techniques

- ▶ Box plot
- ▶ Histogram
- ▶ Multi-vari chart
- ▶ Run chart
- ▶ Pareto chart
- ▶ Scatter plot
- ▶ Stem-and-leaf plot
- ▶ Parallel coordinates
- ▶ Odds ratio
- ▶ Targeted projection pursuit
- ▶ Glyph-based visualization methods such as PhenoPlot and Chernoff faces
- ▶ Projection methods such as grand tour, guided tour and manual tour
- ▶ Interactive versions of these plots
- ▶ Dimensionality reduction

DEMO 1:EXPLORATORY DATA ANALYSIS

https://gits-15.sys.kth.se/amiakh/ID2214/blob/master/practical_session/Exploratory%20Data%20Analysis.ipynb



Why do you need tools and libraries?

Usecase

- ▶ You are hired as a Machine Learning engineer (or a data scientist). Your company is going to start working on a project with fraud detection for a large company.
- ▶ Your company's CTO advises you to use models that can achieve good accuracy and you know that ensemble methods like gradient boosting tree or neural networks are the way to go.
- ▶ When you go and start working with the algorithms you have coded, you realize that they are training very slow with the data you have at hand.



Why do you need tools and libraries?

Usecase

- ▶ You are asked to train and deploy a machine learning model for a customer. In this project you will use a mixture of linear models and neural networks.
- ▶ You start coding your models from scratch, it runs slowly and it has a few bugs. Should you fix it?
- ▶ You go and look at Scikit Learn's implementation and realize that their implementation is more than 600 lines of code even for a linear regression model.

Clock is ticking ...what would you do?

ML projects are not only about modelling

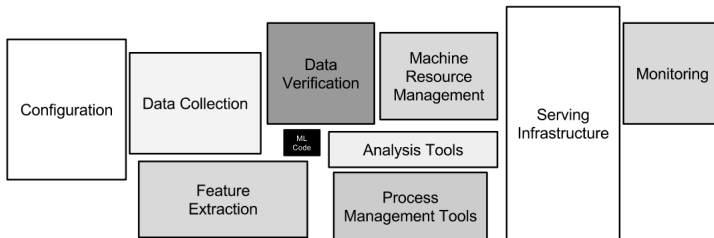


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Real-world ML systems (Sculley et al. [2015])



Advantages of using libraries

- ▶ So far in this course, you have learned to implement:
 - Data pre-processing algorithms
 - Machine Learning models that can predict an outcome from a series of inputs
- ▶ However there are many tools and libraries out there to facilitate your job as data scientist/machine learning engineer.
- ▶ In reality, it is costly to code and maintain your own algorithms, your code needs to
 - Be optimized to run on different operating systems
 - Be optimized under different hardware settings
 - Function with different updates of the programming language you type
 - Carefully consider type checkings, underflow and overflows

Disadvantages of using libraries

- ▶ Your desired features may not have been implemented and might take long to get to the production
- ▶ Some libraries have licenses and regulations when used for production
- ▶ You cannot verify the code-base yourself
- ▶ You cannot follow the change-logs in the library as the development cycles are fast
- ▶ Sometimes new features are not backward compatible

You REALLY need to know your libraries!



Zachary Lipton ✓

@zacharylipton




By default, logistic regression in scikit-learn runs w L2 regularization on and defaulting to magic number $C=1.0$. How many millions of ML/stats/data-mining papers have been written by authors who didn't report (& honestly didn't think they were) using regularization?

6:49 AM · Aug 30, 2019 · [Twitter Web App](#)

247 Retweets **1.3K** Likes

You REALLY need to know your libraries!


 tensorflow / tensorflow

Used by 55.8k Watch 8.6k

[Code](#) [Issues 2,857](#) [Pull requests 219](#) [Actions](#) [Projects 1](#) [Security](#) [Insights](#)

Evaluated expressions of variables differ sometimes using the GPU #2226

Closed alexlee-gk opened this issue on May 5, 2016 · 8 comments

 alexlee-gk commented on May 5, 2016

The evaluated value of the l1 or l2 loss of variables differ sometimes when using the GPU. This seems to happen when the variables are "large".

Even though the difference is not that much in the example below, these differences have resulted in a 98% test accuracy on a network trained on the CPU, but a 10% test accuracy on the same network trained on the GPU.

Environment info

Operating System: Ubuntu 14.04

Installed version of CUDA and cuDNN: 7.5 and 7.0 [cuda_info.txt](#)

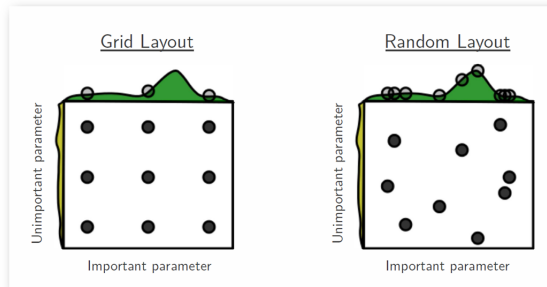
GPU: Titan X



Parameters in Machine Learning

- ▶ Model parameters: Parameters that are learned using an optimization algorithm through training
- ▶ Model hyper-parameters: Parameters you need to set before you can use the model to train, e.g. activation functions in neural networks, depth of a decision tree or regularization parameter in a linear model

Hyperparameter Optimization Techniques



Goodfellow et al. [2016]



Other Modalities of Data

- ▶ The main focus of this course so far has been on data modalities that are presented in numerical and categorical formats
- ▶ This type of data is usually called a tabular data
- ▶ The access to Internet has brought much access to other data modalities such as text and images
- ▶ Text and Image data has characteristics that we call high dimensional data



Text Preprocessing

- ▶ Free-text data is data represented in a human readable language
- ▶ In order to use text in machine learning model, we need to find a vector representation for the data
- ▶ There exists many solutions to this problem, each of which with their strength and shortcomings, i.e. bag of words, n-gram, tf-idf, ...

Bag of Words

- ▶ In this representation, text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.
- ▶ Two sentences: 1) John likes to watch movies. 2) Mary likes movies too.
- ▶ Vocabulary = ['John', 'likes', 'to', 'watch', 'Mary', 'also', 'movies', 'too']
- ▶ BoW1 = [1, 1, 1, 1, 0, 0, 1, 0]
- ▶ BoW2 = [0, 1, 0, 0, 1, 0, 1, 1]

Recommendation algorithms



Image from Google Developers

DEMO 2:PRACTICAL MODELLING

https://gits-15.sys.kth.se/amiakh/ID2214/blob/master/practical_session/Best%20Practice%20in%20Modelling.ipynb

- I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511, 2015.
- Wikipedia. Exploratory data analysis — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Exploratory%20data%20analysis&oldid=926125408>, 2019. [Online; accessed 24-November-2019].