STFC's National CDT Training Event 2019

Report of Contributions

https://indico.cern.ch/e/770601

Contribution ID: **2** Type: **not specified**

Planning your data science skills path

Tuesday 20 November 2018 10:30 (2 hours)

By the end of your PhD you need to build the skills you'll need to compete in industry, academia, or elsewhere. In this session we will plot the path to deciding which skills you'll need, and how to develop them best.

Presenter: TOJEIRO, Rita

Session Classification: Tuesday Morning Training Session

Contribution ID: **3** Type: **not specified**

Git & Github

Tuesday 20 November 2018 10:30 (2 hours)

I will quickly introduce the concept of version control, and distributed version control systems (DVCS), of which the most famous is git. We'll learn how to use git from the command line to create repositories which keep track of our projects (be they code, documents or otherwise). I'll end by describing how to store your git repositories on a remote server, the most famous of which is GitHub. If we have time, we'll get into more advanced topics such as branching.

Presenter: FORGAN, Duncan

Session Classification: Tuesday Morning Training Session

Contribution ID: **4** Type: **not specified**

Getting started with Jupyter notebooks and machine learning

Tuesday 20 November 2018 10:30 (2 hours)

Machine learning is one of the biggest buzzwords in the field of data science and it has many applications within both academia and industry. In this tutorial, Pivigo's community manager and a data scientist will take you through the basics of using Jupyter notebooks and how to get started with machine learning. Jupyter notebooks are a fantastic way to explore data and to conduct experiments on the data. He will be using the titanic dataset on Kaggle.

Presenter: MAHTANI, Deepak

Session Classification: Tuesday Morning Training Session

Contribution ID: **5** Type: **not specified**

Classical unsupervised learning with scikit-learn

Tuesday 20 November 2018 10:30 (2 hours)

Unsupervised learning is a subsection of machine learning where the computer is trained with unlabelled data and must pick out important features in the data on its own. In terms of classical machine learning this boils down to two main approaches: dimensionality reduction and clustering. We will look at these two concepts and apply them practically to a dataset to test how good the computer's intuition is.

Presenter: ARMSTRONG, John

Session Classification: Tuesday Morning Training Session

Contribution ID: **6** Type: **not specified**

Parallel Programming

Tuesday 20 November 2018 14:00 (2 hours)

Getting the most out of modern computer architecture means using parallel programming - making the computer run different pieces of calculations at the same time. In this tutorial we will discuss the various approaches towards this, such as thread- and process-level parallelism, and focus in detail on the MPI approach to the latter.

Presenter: HENTY, David

Session Classification: Tuesday Afternoon Training Session

Contribution ID: **7** Type: **not specified**

Planning your data science skills

Tuesday 20 November 2018 14:00 (2 hours)

By the end of your PhD you need to build the skills you'll need to compete in industry, academia, or elsewhere. In this session we will plot the path to deciding which skills you'll need, and how to develop them best.

Presenter: TOJEIRO, Rita

Session Classification: Tuesday Afternoon Training Session

Contribution ID: **8** Type: **not specified**

Git & Github

Tuesday 20 November 2018 14:00 (2 hours)

I will quickly introduce the concept of version control, and distributed version control systems (DVCS), of which the most famous is git. We'll learn how to use git from the command line to create repositories which keep track of our projects (be they code, documents or otherwise). I'll end by describing how to store your git repositories on a remote server, the most famous of which is GitHub. If we have time, we'll get into more advanced topics such as branching.

Presenter: FORGAN, Duncan

Session Classification: Tuesday Afternoon Training Session

Contribution ID: **9** Type: **not specified**

Data Science in the cloud by DeltaDNA

Tuesday 20 November 2018 14:00 (2 hours)

Cloud computing has moved quickly from simply providing easy access to hardware to supplying easy to use services that eliminate the need for any sys admin knowledge. ML and data science has been one of the key drivers of this, with all the major cloud providers (i.e. AWS, GCP and Azure) offering a suite of tools to allow the construction of everything from data pipelines to ML model fitting to APIs to categorise data in real time and everything in between. In this session I will go over what is available in the cloud for data scientists and talk through some typical cloud tech stacks you will encounter in the world of big data. Finally I will talk about the pros and cons of different cloud technologies and how they apply to different real world applications.

Presenter: ROSEBOOM, Isaac (DeltaDNA)

Session Classification: Tuesday Afternoon Training Session

Contribution ID: **10** Type: **not specified**

Data Science & Machine Learning in e-Commerce by ASOS

Tuesday 20 November 2018 14:00 (2 hours)

Big data in retail and e-commerce (a sort of why, what and how)
- Common machine learning methods in retail and e-commerce (recommender systems, customer lifetime value prediction, automatic product understanding)
- Practical aspects of deploying and using ML in retail and e-commerce (using large distributed computing systems such as Spark for example)
- 'Soft skills'for data scientists: stakeholder management, understanding business value, expectation management.

Presenter: LITTLE, Duncan (ASOS)

Session Classification: Tuesday Afternoon Training Session

Contribution ID: **11** Type: **not specified**

Parallel Programming

Wednesday 21 November 2018 09:00 (2 hours)

Getting the most out of modern computer architecture means using parallel programming - making the computer run different pieces of calculations at the same time. In this tutorial we will discuss the various approaches towards this, such as thread- and process-level parallelism, and focus in detail on the MPI approach to the latter.

Presenter: HENTY, David

Session Classification: Wednesday Morning Training Session

Contribution ID: **12** Type: **not specified**

Neural Networks in TensorFlow

Wednesday 21 November 2018 09:00 (2 hours)

This workshop will briefly introduce supervised deep-learning with neural networks, before walking through an example of constructing and training such networks using keras, a high-level Python interface to TensorFlow. We will train both fully-connected and convolutional neural nets to distinguish hand-written digits from the standard MNIST dataset, and validate the results. I will also discuss the importance of understanding and preparing your training data, and illustrate the benefits of data augmentation.

Presenter: BAMFORD, Steven (Unknown)

Session Classification: Wednesday Morning Training Session

Contribution ID: **13** Type: **not specified**

Data Science & Machine Learning in e-Commerce by ASOS

Wednesday 21 November 2018 09:00 (2 hours)

Big data in retail and e-commerce (a sort of why, what and how)
- Common machine learning methods in retail and e-commerce (recommender systems, customer lifetime value prediction, automatic product understanding)
- Practical aspects of deploying and using ML in retail and e-commerce (using large distributed computing systems such as Spark for example)
- 'Soft skills'for data scientists: stakeholder management, understanding business value, expectation management.

Presenter: LITTLE, Duncan (ASOS)

Session Classification: Wednesday Morning Training Session

Contribution ID: **14** Type: **not specified**

Code profiling & Optimisation

Wednesday 21 November 2018 09:00 (2 hours)

The faster our code runs, the more data we can process. In this tutorial you, will learn how to measure and improve CPU and memory performance in Linux.

Presenter: MARTIN-HAUGH, Stewart (Science and Technology Facilities Council STFC (GB))

Session Classification: Wednesday Morning Training Session

Contribution ID: **17** Type: **not specified**

Neural Networks in TensorFlow

Wednesday 21 November 2018 11:00 (2 hours)

This workshop will briefly introduce supervised deep-learning with neural networks, before walking through an example of constructing and training such networks using keras, a high-level Python interface to TensorFlow. We will train both fully-connected and convolutional neural nets to distinguish hand-written digits from the standard MNIST dataset, and validate the results. I will also discuss the importance of understanding and preparing your training data, and illustrate the benefits of data augmentation.

Presenter: BAMFORD, Steven (Unknown)

Session Classification: Wednesday Morning 2nd Training Session

Contribution ID: **18** Type: **not specified**

Classical unsupervised learning with scikit-learn

Wednesday 21 November 2018 11:00 (2 hours)

Unsupervised learning is a subsection of machine learning where the computer is trained with unlabelled data and must pick out important features in the data on its own. In terms of classical machine learning this boils down to two main approaches: dimensionality reduction and clustering. We will look at these two concepts and apply them practically to a dataset to test how good the computer's intuition is.

Presenter: ARMSTRONG, John

Session Classification: Wednesday Morning 2nd Training Session

Contribution ID: **19** Type: **not specified**

Code profiling & Optimisation

Wednesday 21 November 2018 11:00 (2 hours)

The faster our code runs, the more data we can process. In this tutorial you, will learn how to measure and improve CPU and memory performance in Linux.

Presenter: MARTIN-HAUGH, Stewart (Science and Technology Facilities Council STFC (GB))

Session Classification: Wednesday Morning 2nd Training Session

Contribution ID: **20** Type: **not specified**

GPU Programming

Wednesday 21 November 2018 11:00 (2 hours)

Graphics Processing Units (GPUs) are commonly available computing devices designed to enhancing computer game experiences. The underlying hardware can, however, be exploited to perform general calculations and has given rise to General-Purpose GPU (GPGPU) computing. In this tutorial, I will discuss 1) when it might be advantageous to develop code to run on a GPU, 2) the nuances of GPU hardware that affect the algorithm ported to the GPU, compared specifically with other forms of parallel programming, and 3) examples of GPU programming with CUDA, nVidia's extension to C/C++, highlighting ease and indicating pitfalls. Knowledge of C/C++ is advantageous, but not essential.

Presenter: TITTLEY, Eric

Session Classification: Wednesday Morning 2nd Training Session

Contribution ID: **21** Type: **not specified**

GPU Programming

Wednesday 21 November 2018 09:00 (2 hours)

Graphics Processing Units (GPUs) are commonly available computing devices designed to enhancing computer game experiences. The underlying hardware can, however, be exploited to perform general calculations and has given rise to General-Purpose GPU (GPGPU) computing. In this tutorial, I will discuss 1) when it might be advantageous to develop code to run on a GPU, 2) the nuances of GPU hardware that affect the algorithm ported to the GPU, compared specifically with other forms of parallel programming, and 3) examples of GPU programming with CUDA, nVidia's extension to C/C++, highlighting ease and indicating pitfalls. Knowledge of C/C++ is advantageous, but not essential.

Presenter: TITTLEY, Eric

Session Classification: Wednesday Morning Training Session

Contribution ID: **23** Type: **not specified**

Instructions

This contribution contains the instructions for the various sessions.