CS9D1-15 Foundations of Data Analytics

24/25

Department
Computer Science
Level
Taught Postgraduate Level
Module leader
Greg Watson
Credit value
15
Module duration
1 week
Assessment
100% coursework
Study location
University of Warwick main campus, Coventry

Description

Introductory description

Students will study techniques for how to go from raw data to a deeper understanding of the patterns and structures within the data, to support making predictions and decision making.

Module aims

To understand the foundational skills in data analytics, including preparing and working with data; abstracting and modelling an analytic question; and using tools from statistics, learning and mining to address the question

Outline syllabus

This is an indicative module outline only to give an indication of the sort of topics that may be covered. Actual sessions held may differ.

Data Analytics involves being about to go from raw data to a deeper understanding of the patterns and structures within the data, to support making predictions and decision making. The course will cover a number of topic, including:

Introduction to analytics, case studies - How analytics is used in practice. Examples of successful analytics work from companies such as Google, Facebook, Kaggle, and Netflix. Suggestions for the course project.
Basic tools: command line tools, plotting tools, programming tools - The wide variety of tools available to work with data, including unix/linux command line tools for data manipulation (sorting, counting, reformatting, aggregating, joining); tools such as gnuplot for displaying and visualizing data; advanced programming tools such as Perl and Python for powerful data manipulation.
Statistics: Probability recap, distributions, significance tests - The tools from statistics for understanding distributions and probability (means, variance, tail bounds). Hypothesis testing for determining the significance of an observation.
Database: Data quality, data cleaning, Relational data, SQL, NoSQL - Problems found in realistic data: errors, missing values, lack of consistency, and techniques for addressing them. The relational data model, and the SQL language for expressing queries. The NoSQL movement, and the systems evolving around it.
Regression: linear regression, least squares, logistic regression - Predicting new data values via regression models. Simple linear regression over low dimensional data, regression for higher dimensional data via least squares optimization, logistic regression for categoric data.
Matrices: Linear Algebra, PCA - Matrices to represent relations between data, and necessary linear algerbraic operations on matrices. Application to the Netflix prize.
Clustering: hierarchical, k-means, k-center - Finding clusters in data via different approaches. Choosing distance metrics. Different clustering approaches: hierarchical agglomerative clustering, k-means (Lloyd's algorithm), k-center approximations. Relative merits of each method.
Classification: Trees, NB, Support Vector Machines - Building models to classify new data instances. Decision tree approaches and Naive Bayes classifiers. The Weka toolkit.
Data Structures: Data structures to scale analytics to big data and data streams.
Artificial Neural Networks: The use of artificial neural networks for tasks such as regression and classification, plus the relationship between artificial neural networks and deep learning.
Data Sharing: Privacy, Anonymization, Risks - The ethics and risks of sharing data on individuals. Technologies for anonymizing data: k-anonymity, and differential privacy.

Learning outcomes

By the end of the module, students should be able to:

- Identify issues with scaling analytics to large data sets, and use appropriate techniques (NoSQL systems, data structures) to scale up the computation.
- Appreciate the need for privacy, identify privacy risks in releasing information, and design techniques to mediate these risks.
- Understand the principles and purposes of data analytics, and articulate the different dimensions of the area.
- Work with and manipulate a data set to extract statistics and features, coping with missing and dirty data.
- Apply basic data mining machine learning techniques to build a classifier or regression model, and predict values for new examples.

Indicative reading list

Recommended Text:
Data Mining: Concepts and Techniques. Jiawei Han, Michelle Kanber, Jian Pei. Morgan Kaufman, 2011
Additional Reading:
Data Manipulation with R. Phil Spector. Springer, 2008
Machine Learning. Thom Mitchell. McGraw Hill, 1997
Database Systems: An Application-oriented Approach, Introductory Version. Michael Kifer, Arthur Bernstein, Philip Lewis. Addison Wesley, 2004
The Works: Anatomy of a City. Kate Ascher. Penguin, 2012

Subject specific skills

Working with data;
abstracting and modelling;

Transferable skills

Communication skills;
Problem solving.

Study

Study time

Type	Required
Lectures	20 sessions of 1 hour (17%)
Practical classes	10 sessions of 1 hour (8%)
Work-based learning	30 sessions of 1 hour (25%)
Online learning (independent)	30 sessions of 1 hour (25%)
Assessment	30 hours (25%)
Total	120 hours

Private study description

No private study requirements defined for this module.

Costs

No further costs have been identified for this module.

Assessment

You must pass all assessment components to pass the module.

Assessment group A1

	Weighting	Study time
Data Analysis Project	100%	30 hours

This assessment is worth more than 3 CATS and is, therefore, ineligible for self-certification.

Feedback on assessment

Written feedback will be provided by the module organiser.

Availability

There is currently no information about the courses for which this module is core or optional.